Design and analysis of biomedical studies

Hansen, Merete Kjær

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Design and analysis of biomedical studies

Merete Kjær Hansen

Kongens Lyngby 2014
PhD-2014-343

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Matematiktorvet, building 303B,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3351
compute@compute.dtu.dk
www.compute.dtu.dk
PhD-2014-343

Summary

Biomedicine is a field that has great influence on the majority of mankind. The constant development has considerably changed our way of life during the last centuries. This has been achieved through the dedication of biomedical researchers along with the tremendous resources that over time have been allocated this field. It is utterly important to utilize these resources responsibly and efficiently by constantly striving to ensure high-quality biomedical studies. This involves the use of a sound statistical methodology regarding both the design and analysis of biomedical studies. The focus of this project is on statistical aspects that arise within the field of biomedicine.

Two types of errors are frequently accentuated within the framework of statistics, namely type I and type II errors. Type I errors occur when a null hypothesis erroneously is rejected. An acceptable type I error rate is specified prior to conducting the statistical analysis. However, all statistical models make assumptions and if violated the actual type I error rate may deviate from the pre-specified type I error rate. Type II errors occur when we fail to reject a false null hypothesis. On contrary to the type I error rate, the type II error rate is not explicitly specified during the statistical analysis and this entails that assessment of the type II error rate in practice is at risk of being neglected altogether. Concerns regarding type I errors, type II errors and adherence (or lack thereof) to model assumptions for biomedical studies are a recurring theme in this thesis.

Data collected in some biomedical studies are positively skewed; hence methods relying on the normal distribution are not directly applicable. We investigated how data from one of these studies are suitably analyzed. We extracted 23 different summary statistics from data gathered from eleven studies. The degree

of adherence to the model assumptions evaluated for each of these summary statistics form basis for our conclusions.

Hierarchically structured data are frequently encountered in biomedical studies. For one type of studies entailing such data we have conducted a literature study strongly indicating that this structure commonly is neglected in the statistical analysis. Based on this closed-form expressions for the approximate type I error rate are formulated. The type I error rates are assessed for a number of factor combinations as they appear in practice and in all cases the type I error rates are demonstrated to be severely inflated.

Prior to conducting a study it is important to perform power and sample size determinations to ensure that reliable conclusions can be drawn from the statistical analysis. We have formulated closed-form expressions for the statistical power of studies with a hierarchical structure to guide biomedical researchers designing future studies of this type.

Upon model fitting it is important to examine if the model assumptions are met to avoid that spurious conclusions are drawn. While the range of diagnostic methods is extensive for models assuming a normal response it is generally more limited for non-normal models. An R package providing diagnostic tools suitable for examining the validity of binomial regression models have been developed. The binomTools package is publicly available at the CRAN repository.

Resumé

Biomedicin er et område der har stor indflydelse på størstedelen af menneskeheden. Den vedvarende udvikling har gennem de sidste århundreder ændret vores levevis væsentligt. Dette er opnået gennem biomedicinske forskeres dedikerede indsats og de enorme ressourcer som gennem tiden er afsat til dette område. Det er vigtigt at udnytte disse ressourcer ansvarligt og effektivt ved konstant at bestræbe sig på at sikre biomedicinske studier af høj kvalitet. Dette indebærer at passende statistiske metoder anvendes både i forhold til design og analyse af biomedicinske studier. Dette projekt har fokus på statistiske aspekter der opstår indenfor biomedicin.

To typer af fejl fremhæves indenfor statistik, hvilket er type I og type II fejl. Type I fejl optræder når nulhypotesen fejlagtigt forkastes. En acceptabel type I fejlrate specificeres før den statistiske analyse udføres. Alle statistiske modeller bygger dog på antagelser og den faktiske type I fejlrate kan afvige fra den præspecificerede type I fejlrate, hvis modellens antagelser ikke er opfyldt. Type II fejl optræder når vi undlader at forkaste en falsk nulhypotese. I modsætning til type I fejlraten bliver type II fejlraten ikke eksplicit specificeret i den statistiske analyse, hvilket medfører risiko for at type II fejlraten ikke tages i betragtning. Overvejelser omkring type I fejl, type II fejl og opfyldelse (eller mangel på samme) af modelantagelser i relation til biomedicinske studier er et gennemgående tema i denne afhandling.

I nogle biomedicinske studier er de indsamlede data positivt skævt fordelt, og metoder der beror på normalfordelingen kan derfor ikke anvendes direkte. For et af disse studier har vi undersøgt hvordan data kan analyseres. Vi har uddraget 23 forskellige nøgletal fra data samlet fra elleve studier. Vores konklusioner baserer sig på graden af modelantagelsernes opfyldelse.

Data med en hierarkisk struktur optræder hyppigt i biomedicinske studier. For et af disse studier har vi udført et litteraturstudie der stærkt indikerer at denne struktur ofte negligeres i den statistiske analyse. Baseret herpå har vi formuleret et lukket udtryk for den approksimative type I fejlrate. Type I fejlraten for forskellige faktorkombinationer som de optræder i praksis er i alle tilfælde påvist at være stærkt forøget.

Før et studie udføres er det vigtigt at beregne styrke og stikprøvestørrelse for at sikre at pålidelige konklusioner senere kan drages på baggrund af den statistiske analyse. Vi har formuleret et lukket udtryk for den statistiske styrke af studier med hierarkisk struktur med det formål at guide biomedicinske forskere når fremtidige studier af denne type skal designes.

Når en model er opstillet er det vigtigt at undersøge om modelantagelserne er opfyldt for at undgå at fejlagtige konklusioner drages. Omfanget af diagnostiske metoder for modeller baseret på normalfordelingsantagelsen er omfattende, hvorimod udvalget for andre typer af modeller generelt er mere begrænset. En R pakke med diagnostiske redskaber til at undersøge modelvaliditet af binomiale regressionsmodeller er udviklet. Pakken binomTools er offentligt tilgængeligt på CRAN repositoriet.

Preface

This thesis was prepared at the Technical University of Denmark (DTU), Department of Applied Mathematics and Computer Science (DTU Compute), Section of Statistics and Data Analysis in partial fulfillment of the requirements for acquiring the Ph.D. degree in Applied Mathematical Statistics. The project was funded by the Technical University of Denmark and was supervised by Murat Kulahci who took over from Klaus Kaae Andersen.

The thesis deals with different statistical aspects of the design and analysis of biomedical studies. More specifically, the work herein relates to issues that are encountered in biological, medical and pharmaceutical studies. These issues however are not limited to these types of studies and are of pertinence in various fields.

The thesis consists of four research papers, one technical report and one R package that is documented by its reference manual. An introductory part provides an overview of the thesis, background information and a summary of the results.

Lyngby, July 2014

Merete Kjær Hansen

List of contributions

Manuscripts

[A] **Hansen, M. K.** and Kulahci, M. (2014). Assessment of the type I error rate when ignoring the hierarchical structure of *in vivo* Comet assay data. In Peter Linde (ed.), *Symposium i Anvendt Statistik* [Symposium in Applied Statistics], University of Copenhagen, p. 83–92, non peer-reviewed.

[B] **Hansen, M. K.** and Kulahci, M. (2014). *The type I error rate for in vivo Comet assay data when the hierarchical structure is disregarded.* DTU Compute Technical Report No. 9. Department of Applied Mathematics and Computer Science, Technical University of Denmark.

[C] **Hansen, M. K.** and Kulahci, M. (2014). On the Type I Error Rate When the Hierarchical Structure of Data Is Ignored. *The American Statistician*, submitted.

[D] **Hansen, M. K.**, Sharma, A. K., Dybdahl, M., Boberg, J. and Kulahci, M. (2014). *In vivo* Comet assay - statistical analysis and power calculations of mice testicular cells. *Mutation Research - Genetic Toxicology and Environmental Mutagenesis*, *774*, 29-40.

[E] Kjeldsen, L. J., Birkholm, T., Fischer, H., Graabæk, T., **Hansen, M. K.**, Kibsdal, K. P., Ravn-Nielsen, L. and Truelshøj, T. H. (2014). A national drug related problems database - evaluation of use in practice, reliability and reproducibility. *International Journal of Clinical Pharmacy*, *36*, 742-749.

[F] Christensen, R. H. B. and **Hansen, M. K.** (2012). binomTools: Performing diagnostics on binomial regression models. R package version 1.0-1. http://CRAN.R-project.org/package=binomTools/

Presentations

Invited

- Workshop: Standardisation of the comet assay - high throughput systems and automated scoring. *The National Research Centre for the Working Environment, April 8-9, 2014, Copenhagen, Denmark.*

Contributed

- Presenting the paper: "Assessment of the type I error rate when ignoring the hierarchical structure of *in vivo* Comet assay data". *36. Symposium i Anvendt Statistik, January 27-29 2014, Department of Economics, University of Copenhagen.*
- The binomTools package: Performing diagnostics on binomial regression models. *UseR!: The R User Conference 2011, August 16-18 2011, University of Warwick, Coventry, UK.*

Posters

- Sharma A.K., Mortensen A., Wedebye E.B., Dybdahl M., **Hansen M.K.**, Kulahci M. Investigation of Germ Cell Genotoxicity by Using the in Vivo Comet Assay in Testis Cells of Mice. *The XIII International Congress of Toxicology, Seoul, Korea, June 30-July 4 2013.*

Acknowledgements

First of all I wish to thank my supervisor Murat Kulahci for providing guidance and inspiration along the course of this PhD project. At the time where my former supervisor was presented new opportunities you willingly took over the supervision for which I am very grateful. I have benefitted appreciably from your excellence not only in statistics but also regarding your superior communication skills that truly are prodigious.

I also would like to thank my former supervisor Klaus Kaae Andersen from The Danish Cancer Society. You were the first to introduce me to the awesomeness of statistics in your inspirational lectures and you later facilitated the transition from biotechnology to statistics through your enthusiastic supervision and encouragement. In the same regard I would like to thank Henrik Spliid that together with Klaus Kaae Andersen provided me the possibility to participate in interesting projects and to develop my statistical understanding during my employment at IMM's Statistical Consultancy Center (ISCC).

I would like to thank Anoop Kumar Sharma from National Food Institute, DTU for providing insight, ideas and data from Comet assay studies. You have been very helpful and our discussions on various topics related to Comet assay studies have been really rewarding.

While it is well-known that life as a PhD student rely on large quantities of coffee there is no guarantee that it can be enjoyed in rewarding company. It has been an utmost pleasure to share the office, coffee and numerous croissants with Rune Haubo B. Christensen. More importantly, the coffee chats often converted into statistical discussions from which I have learned many tips, tricks, dos and

don'ts over the past years. I would like to thank the always cheerful David Meisch for taking over as a highly valued office mate.

Most of all I would like to thank my husband Martin and our children Olivia and Gustav. You embrace me with love and have patiently supported me through all my ups and downs during this project.

Contents

Chapter 1

Introduction

Biomedicine is an eminent field that relates to the majority of mankind. It does so directly when e.g. individual tests are run to clarify the condition of a patient and indirectly when e.g. treatment of specific conditions can be commended based on knowledge gathered through biomedical research. Usually, the achievements of biomedical research are not pertinent in peoples mind, yet it is an area that has an impact on everyone. In the extreme it makes the difference between life and dead, such as when a child is affected by pneumonia. Through the accomplishments of biomedicine it is nowadays generally possible to diagnose and successfully treat such a patient as opposed to a century ago.

One great biomedical victory is the discovery of penicillin by Alexander Fleming in 1928. This attainment is truly peculiar as one of the key elements was in fact untidiness. After spending a few weeks on vacation, Fleming returned to his lab where he was met by quite a disagreeable smell. In the clean-up process he found some forgotten petri dishes with *Staphylococcus aureus* that accidentally had been contaminated with the fungus *Penicillium notatum*. He noticed that no bacterial colonies appeared close to the fungus, which led him to speculate whether it could be associated with some sort of antibacterial activity. Further research to elucidate this theory was initiated, but it was not until 1944 that a mass production of penicillin was possible. The popularity was however immediate and penicillin have been suggested to change the course of World War II in favor of the allies due to the recovery of innumerable soldiers. Fleming

was smiled on by fortune but that alone does not explain his success. Also vigilance and dedication were crucial qualities that Fleming possessed. In general, considerable knowledge has in the course of centuries been attained from biomedical studies through dedication, creativity, discipline, luck, innovation and enormous resources.

Considering the significance of biomedical studies and the massive resources they are allocated, it is important to ensure a high quality and to use these resources in an efficient and responsible manner. Upon collection of data a continual challenge is to judge whether the observations reflect a true effect of interest or stem from pure randomness. Statistics is useful in mitigating these concerns. This is increasingly recognized by research communities across a variety of fields and some sort of statistical treatment is often a prerequisite for publications to be accepted. On the other hand, a persistent characteristic of applied statistics is its lack of self-sufficiency. Applied statistics is in general applied to data that naturally associates with and are collected within the framework of other research areas. Collaboration between statistical practice and empirical biomedical sciences can thus be mutually beneficial, which seems to be increasingly recognized.

However, design of experiments and the statistical analysis of data from empirical studies are still not always conducted in collaboration with a statistician, data analyst or the like. There may be many reasons for this, including common practice, lack of resources, time constraints, unawareness of a flawed methodology, difficulties in the communication between biomedical researchers and statisticians etc. This is unproblematic whenever the complexity level of the statistical issues does not surpass the statistical qualifications of the relevant researchers. In some cases, though, inappropriate methods are inadvertently applied, which may lead to inefficient designs and analyses and even that flawed inferences are made.

The misuse of statistics in medical studies has been addressed repeatedly by several authors (Yates and Healy, 1964; Altman, 1981; Festing et al., 2002; Gardenier and Resnik, 2002; Baccaglini et al., 2010). According to Strasak et al. (2007): "Standards in the use of statistics in medical research are generally low. A growing body of literature points to persistent statistical errors, flaws and deficiencies in most medical journals". They state this in a comprehensive review summarizing 45 papers that nearly all are concerned with the inappropriate use of statistically related aspects that too often occur within biomedical research. The review classify the statistical misuse into different stages of a scientific study, which include the design, data analysis, documentation, presentation and the interpretation of the results. To ensure a sufficient quality of biomedical studies the statistical practice in all of these stages is rather essential, although it seems to be infringed too often. The problem may in reality be

Table 1.1: The four possible outcomes of a hypothesis test. The probability of the outcomes are seen in parentheses

	reject H_0	fail to reject H_0
H_0 is true	type I error (α)	correct failure of rejection
H_0 is false	correct rejection (power)	type II error (β)

even greater as some kind of misuses are not evident based on the description of the statistical methods.

Within a statistical framework a null hypothesis (H_0) is usually defined. Based on the collected data we either reject or fail to reject this hypothesis. Consequently, there are basically two types or errors[1] that can be committed: a type I error, where the null hypothesis erroneously is rejected and a type II error, where the null hypothesis incorrectly is not rejected (see Table 1.1). Depending on the context one type of these errors may be considered more serious than the other. In many fields the type I error is most often considered the more serious and this type of error is therefore more strictly controlled than the type II error. In practice, this is done by fixing the type I error rate (denoted α) at a prespecified level; often at 0.05 or 0.01. Secondarily, the type II error rate (denoted β) is sought minimized by choosing the uniformly most powerful test (if it exists), adjusting the sample size etc.

An example of committing a type I error is when a drug is deemed to be effective when it in fact is not. Within biomedical research it is possible that follow-up studies will be conducted on grounds of a significant finding and type I errors are therefore problematic from an ethical and financial perspective. The occurrence of type I errors are impossible to avoid but it is crucial to carefully control the error rate and ensure that they do not surpass the appointed acceptance level.

A type II error can be seen as the opposite of a type I error rate, that is, when e.g. a drug is concluded to be ineffective although it is not. In many fields it is common to aim for the type II error rate to be lower than 10% or 20% (i.e. the statistical power is above 80% or 90%). However, in practice it is often not considered and the magnitude of the type II error rate are in these cases unknown. This can be very problematic in different settings. For instance, biomedical studies conducted on humans or animals are associated with great ethical concerns and the workload and expenses are often considerable. Therefore, it is important to utilize all resources fully; using too many subjects is obviously a waste but it may be of an even greater concern to use too few subjects as the

[1] Additional types of errors (type III and type 0) have been defined but are not considered in this thesis

risk of committing a type II error in that case becomes intolerable.

When a study suffers from an inappropriate statistical methodology it likely affects the type I and/or the type II error rate. Depending on the specific type of misuse the error types can be severely inflated. To avoid these repercussions it is thus crucial to be aware of the assumptions underlying the statistical method. This applies both to the choice of method as to how the results are interpreted.

Another important concern relates to the documentation of the statistical methodology that in some research papers is very limited and inaccurate. One example is when a method is briefly stated such as ANOVA without mentioning which variables are included and without any explicitly defined model. Since ANOVA is the term for a collection of statistical methods it is technically impossible to deduce exactly how exactly the analysis is performed. This is problematic as the interpretation of the results is closely linked to the method that is used. In addition to this, the documentation in some research papers within certain biomedical research fields can be surprisingly verbatim. This could indicate that some authors are highly inspired by the wording in other publications, which again may imply that the authors do not have a deep insight into the field of statistics. While this is completely understandable, it however entails the pitfall that the documentation and the statistical practice are not in full agreement. This also interferes severely with a proper interpretation of the results.

1.1 Aim of the thesis

Biomedicine is a broad field not unanimously defined. In this thesis we consider studies related to health care and public health, encompassing studies that are within the fields of biology, medicine, genetics, pharmacology and related areas. In particular, this thesis deals with the statistical design and analysis of biomedical studies.

There is no doubt that great effort continuously is made by biomedical and statistical researchers in order to conduct biomedical studies soundly and scientifically. Still, there seems to be a need for further work and research in the area that constitutes the interface between statistics and biomedicine.

The aim of this thesis is to contribute to bridging the gap between biomedical sciences and statistical practice. This has been approached by addressing specific issues that appear in the literature. The focus includes communicating statistical principles and the current research findings to biomedical researchers.

Therefore, derivations are provided when it is considered relevant, while at other times the results are provided in terms of figures, tables and discussions in line with common practice within the specific research area that constitutes the target audience.

1.2 Thesis outline

This thesis consists of seven chapters that provide an introduction to the appended papers and the R package. A summary of the papers are included in the relevant chapters. The papers and a reference manual for the R package are listed in Appendix A-F.

A considerable part of this work concerns data from Comet assay studies and an introduction to these types of studies and the resulting data is given in Chapter 2. Chapter 3 concerns the implications when a hierarchical structure in data is disregarded in the statistical analysis. This is a violation of the critical assumption of independence, which impose severe inflation of the type I error rate. In Chapter 4 the type II error rate and closely related statistical power are studied for hierarchical models and Comet assay studies in particular. As a part of this it is examined how to suitable analyze Comet assay data to comply with relevant model assumptions. Chapter 5 deals with the statistical analysis of agreement data. Much confusion shrouds the analysis of this type of data due to the concurrent popularity and criticism of the kappa statistic that in some fields has become the *de facto* standard in the analysis of agreement data. Chapter 6 gives a presentation of the R package binomTools. This software contains an implementation of a range of diagnostic methods for binomial regression models. Concluding remarks are given in Chapter 7.

Chapter 2

Comet assay studies

Damage to our DNA occurs continuously due to both endogenous (e.g. metabolic processes) and exogenous (e.g. environmental agents) factors. DNA repair mechanisms are effective and constantly active, but some damages are irreparable. Accumulation of damages to our DNA may eventually become hazardous, as it among other things may lead to unregulated cell division and tumors may evolve.

The damages materialize in different ways, and some of them appear as breaks in the DNA strands. The Comet (or Single Cell Gel Electrophoresis) assay is a powerful technique for examining this type of damage in individual cells by quantifying the DNA strand breaks (Lovell and Omori, 2008). The applications of the Comet assay are versatile (Collins et al., 1997; Collins, 2004) and include assessment of the safety of potential new drugs and of possible genotoxicity induced by contaminants. A considerable part of this thesis deals with aspects regarding the design and analysis of Comet assay studies and an introduction to these types of studies are given in the current chapter.

The general principle of the Comet Assay is as follows: a sample of cells is embedded in agarose gel on a slide. The cells are lysed (disrupting the cell membrane resulting in release of the cell content), such that the main remains are the DNA, which naturally are in a supercoiled state (the shape of the DNA caused by winding of the DNA strands). The slide is subjected to electrophoresis

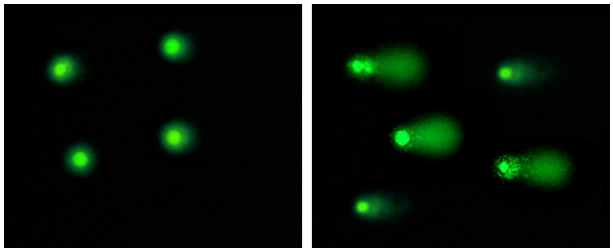

Figure 2.1: Comet assay micrographs of untreated cells to the left and cells that are treated with a DNA damaging agent to the right. Although the formal name of the assay is Single Cell Gel Electrophoresis, it is more commonly referred to as the Comet assay due to the comet-like appearance of damaged cells. Reprinted with kind permission from Cell Biolabs.

thereby exposing the negatively charged DNA to an electric field. The resulting structure observed by fluorescent microscopy often resembles a comet as illustrated in Figure 2.1 (Collins, 2004; Kumaravel and Jha, 2006).

The hypothesized underlying mechanism of the comet-shaped formation is that breaks in the DNA strands implicate a relaxation of the DNA supercoil structure and that damaged DNA are more free to migrate than undamaged DNA. The shape of the comet is thus interpreted as a measure of the number of DNA strand breaks (Collins, 2004; Olive and Banáth, 2006).

2.1 Comet assay data

Several measures seeking to quantify the shape of the obtained comet and thereby the degree of DNA damage have been suggested. The most commonly used are the % tail DNA (also called the tail intensity, TI), the tail length and the Olive tail moment (tail length × % tail DNA). The % tail DNA has gradually become recognized as the most suitable end point. Among other things this is due to its comparability across studies and that it, up to a certain threshold, has been shown to be linearly related to break frequency (Collins et al., 1996; Collins, 2004; Kumaravel and Jha, 2006; Lovell and Omori, 2008). In the present study all data are quantified according to the % tail DNA end point.

In most studies 50 or 100 cells are scored on each slide and the shape of the individual electrophoresed cells are fairly distinct. As illustrated in Figure 2.2 the % tail DNA distribution is strongly positively skewed and roughly takes values between 0 and 80%. When cells are exposed to a genotoxic agent causing

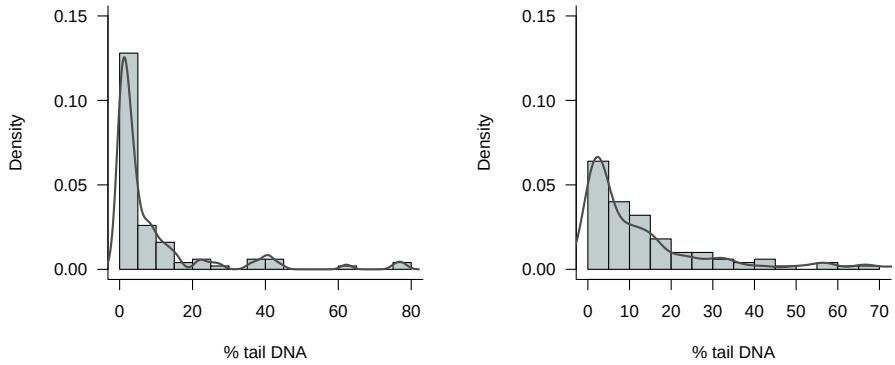

Figure 2.2: Example of data emerging from a Comet assay study. To the left is a % tail DNA distribution summarizing the cells scored on one slide belonging to a vehicle group. To the right is the % tail DNA distribution appertaining to a slide from a positive control group. The animals in the positive control group have been exposed to a DNA damaging agent known to induce DNA strand breaks, which usually inflicts additional skewness to the % tail DNA distribution.

DNA strand breaks, the response level is in general expected to increase due to the relaxation of the DNA supercoil structure. This will in turn impose additional skewness to the observed distribution of the % tail DNA.

The Comet assay experiments generating the present data are conducted as follows: animals are randomly assigned to one of four different treatment groups, i.e. one vehicle group and three groups administered increasing doses of the compound of interest. There are five animals in each group. One or more sample of cells are collected from each animal, put on slides and processed as described above yielding a response of % tail DNA for each cell. This setup imposes a hierarchical structure of data, that is, slide is nested within animal that in turn is nested within treatment. This structure is illustrated in Figure 2.3. Often the interest lies in the assessment of the genotoxic effect potentially induced by the specific doses of the compound that is tested. The specific animals used in the study are not of particular interest but merely act as representatives of the general population of that species.

2.2 Statistical analysis of Comet assay data

There are various approaches regarding the statistical analysis of Comet assay data. Due to the skewed nature it has been suggested to model this type of data by means of the Weibull distribution (Ejchart and Sadlej-Sosnowska, 2003;

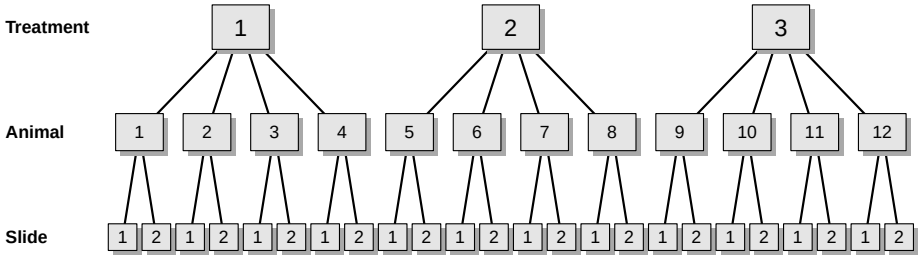

Figure 2.3: Outline of the design commonly used in Comet assay studies. This example shows three treatment groups, four animals per treatment and two slides per animal. For each slide a number of cells are scored, usually in the range of 50-100 cells.

Verde et al., 2006), however it seems that only statistical methods relying on the normal distribution are used in practice (see e.g. the literature study in Paper B for further details). As Comet assay data are hierarchically structured it is crucial to reflect this in the statistical analysis. This can be attained by appropriately summarizing data or by accommodating the structure in the model. In the following, three related statistical models valid for fitting Comet assay data are presented, serving as a reference for the current and the succeeding two chapters. When data are balanced and normally distributed all three methods are equivalent. Due to the assumption of normally distributed data it may be requisite to transform data prior to the statistical modeling.

2.2.1 Using raw cell scores as the response

When the raw cell scores are used as the response the hierarchical structure of data and the randomly selected animals should be properly accounted for. This can be done by employing a linear mixed-effects model with treatment as a fixed effect and animal and slide as random effects. Animal is nested within treatment and slide is nested within animal:

$$Y_{ijkl} = \mu + \tau_i + \beta_{(i)j} + \gamma_{(ij)k} + \varepsilon_{(ijk)l} \tag{2.1}$$

where

$$i = 1, ..., a,\ j = 1, ..., b,\ k = 1, ..., c,\ l = 1, ..., n,$$
$$\beta_{(i)j} \sim N(0, \sigma^2_\beta), \quad \gamma_{(ij)k} \sim N(0, \sigma^2_\gamma), \quad \varepsilon_{(ijk)l} \sim N(0, \sigma^2).$$

Y_{ijkl} is the $ijkl$th observation (one score for each cell) and μ and τ_i are the fixed effects for the intercept and treatment, respectively. $\beta_{(i)j}$ is the random effect

of the jth animal nested within the ith treatment, $\gamma_{(ij)k}$ is the random effect of the kth slide nested within the ith treatment and jth animal and $\varepsilon_{(ijk)l}$ is the within-group error. The parentheses in the subscripts indicate the nesting structure with the parent level(s) given inside the parentheses. See Montgomery (2005) for a more elaborate exposition of the linear mixed-effects model with nested effects.

2.2.2 Summarizing the response for each slide

Another way to analyze data is to summarize the % tail DNA distribution for each slide into a single summary statistic and use this measure in the subsequent analysis. Due to the hierarchical structure of data and the randomly selected animals a suitable analysis of the summarized data is a linear mixed-effects model with treatment as a fixed effect and animal as a random effect and with animal nested within treatment:

$$Y_{ijk} = \mu + \tau_i + \beta_{(i)j} + \varepsilon_{(ij)k} \tag{2.2}$$

where

$$i = 1, ..., a,\ j = 1, ..., b,\ k = 1, ..., n,$$

$$\beta_{(i)j} \sim N(0, \sigma^2_\beta), \quad \varepsilon_{(ij)k} \sim N(0, \sigma^2).$$

Y_{ijk} is the summary statistic of interest calculated for each slide and μ and τ_i are the fixed effects for the intercept and treatment, respectively. $\beta_{(i)j}$ is the random effect of the jth animal nested within the ith treatment and $\varepsilon_{(ij)k}$ is the within-group error.

2.2.3 Summarizing the response for each animal

A third option is to calculate a summary statistic for each animal and use this as the response. A suitable model is the fixed-effects model with treatment as a fixed effect:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \tag{2.3}$$

where

$$i = 1, ..., a,\ j = 1, ..., n,$$

$$\varepsilon_{ij} \sim N(0, \sigma^2).$$

Y_{ij} is the summary statistic of interest calculated for each animal, μ and τ_i are fixed effects for the intercept and treatment, respectively, and ε_{ij} is the within-group error.

Model (2.3) is often referred to as a one-way ANOVA. As will be discussed in Chapter 3 it appears that rather than fitting this model to the cell scores summarized for each animal, it is sometimes applied to the raw cell scores or the statistics summarized for each slide. In case of the latter model (2.3) rather should be formulated as

$$Y_{ij'} = \mu + \tau_i + \varepsilon_{ij'} \tag{2.4}$$

where

$$i = 1, ..., a,\ j' = 1, ..., bn,$$

$$\varepsilon_{ij'} \sim N(0, \sigma'^2).$$

$Y_{ij'}$ is the summary statistic for each slide, μ and τ_i are fixed effects for the intercept and treatment, respectively, and $\varepsilon_{ij'}$ is the assumed within-group error.

2.2.4 Model choice for Comet assay data

Whenever the normality assumption is met, possibly by transformation, model (2.1)-(2.3) are all valid approaches. Yet, (2.2) has throughout this thesis served as the model of choice both in regard to the analysis of Comet assay data and the power and sample size determinations that are provided in Paper D. Given the hierarchical structure of data it is natural to fit a model that reflects this structure. A hierarchical model suitably handles missing observations and information about the animal-to-animal variation compared to the residual variation is readily available. Conversely, the cells are often scored by imaging software and in some versions, if not all, it is possible to extract a summary statistic such as the mean or median directly without having to deal with the raw cell scores. This simplifies the data handling considerably and it is not uncommon practice to make use of summary statistics in the analysis of Comet assay data (see e.g. Paper B). Also, the use of summary statistics in the statistical analysis to some extend stabilizes the distribution of data thereby alleviating assumptional concerns. Altogether, we find model (2.2) to conveniently balance these different concerns.

2.3 Which summary statistic?

A natural question that arises is which summary statistic to employ. Different summary statistics have been proposed, including the mean (Lovell et al., 1999; Wiklund and Agurell, 2003; Bright et al., 2011), median (Lovell et al., 1999; Wiklund and Agurell, 2003; Duez et al., 2003; Bright et al., 2011), 75th percentile (Lovell et al., 1999; Duez et al., 2003) and the 90th percentile (Wiklund and Agurell, 2003; Duez et al., 2003). Also, to comply with the skewness of the within-sample distributions it has been suggested to log-transform the raw data prior to the summary calculations (Lovell and Omori, 2008). Although a few studies specifically address this issue (Wiklund and Agurell, 2003; Duez et al., 2003), they are concerned with different end points than the % tail DNA, and in practice there is no consensus as to which statistic most appropriately summarizes data.

One aim of Paper D is to identify the statistic that most suitably summarizes the % tail DNA distribution for each slide. This assessment is based on data gathered from 11 separate Comet assay studies. Details about these studies are found in Paper D.

The eligibility of a range of summary statistics is assessed: the mean, median (50th), 55th, 60th, 65th, 70th, 75th, 80th, 85th, 90th and 95th percentile of both the raw data as well as data subjected to the natural logarithm are calculated. Also, the log-transformed mean of the raw data is calculated and will be referred to as log(mean). In total, 23 candidate summary statistics are extracted.

Given the importance of model assumptions, the criteria for selecting a summary statistic are founded herein. As the data are fitted a linear mixed-effects model as defined in (2.2), the criteria are accordingly established in decreasing order of significance: 1) variance homogeneity, 2) normality and 3) uncertainty of estimates. The first two criteria are directly derived from the assumptions underlying the linear mixed-effects model (Pinheiro and Bates, 2000). The variance homogeneity assumption may be violated in two distinct ways; 1a) the variance does not remain constant over the range of estimated mean values and 1b) the variance does not remain constant across dose groups. Accordingly, the assessment of the variance homogeneity assumption is two-fold.

For each study and summary statistic, model (2.2) is fitted and the standardized residuals are calculated. To assess criterion 1a) a linear model is fitted regressing the square root of the absolute value of the standardized residuals on the fitted values from model (2.2) and the p-values of the slopes are extracted. Criterion 1b) is assessed by applying Brown-Forsythe's test (also known as the modified Levenes test) (Brown and Forsythe, 1974), which is robust to possible departures

from an underlying normal distribution, to the standardized residuals and the p-values are calculated. The normality assumption addressed in criterion 2) is evaluated by means of Shapiro-Wilk's test (Shapiro and Wilk, 1965), which is applied to the standardized residuals and the p-values are obtained. It should be stressed that while p-values often serve as a mean to decide whether a hypothesis should be rejected or not, the p-values in this setting are solely intended as a measure of the relative performance among the different candidates.

Criterion 4) concerns the uncertainty of the estimates. The within-slide distributions are positively skewed and bear some resemblance with a log-normal distribution. The asymptotic variance of the percentiles of a given distribution is found as

$$\mathrm{Var}(\pi_p) = \frac{p(1-p)}{n(f(F^{-1}(p)))^2}, \tag{2.5}$$

where π_p is the pth percentile, f is the relevant probability density function and $F^{-1}(p)$ is the corresponding quantile function (Mosteller, 1946; Cox and Hinkley, 1974). The variance of the mean of the log-normal distribution is

$$\mathrm{Var}(E(Y)) = \frac{e^{2\mu+\sigma^2}(e^{\sigma^2}-1)}{n} \tag{2.6}$$

(Kotz et al., 2000), where $Y \sim \ln\mathcal{N}(\mu, \sigma^2)$. For the normal distribution the variance of the mean is given as

$$\mathrm{Var}(\mu) = \frac{\sigma^2}{n} \tag{2.7}$$

(Kotz et al., 2000). The asymptotic variances are calculated for the log-normal distribution (resembling the raw within-slide distribution) and the normal distribution (resembling the log-transformed within-slide distribution) with parameter values $\mu = 2$, $\sigma^2 = 2$ and $n = 100$. The variance of log(mean) is not provided since it does not readily compare to the variance of the remaining summary statistics.

The p-values extracted according to criteria 1 and 2 are shown in Figure 2.4. The p-values related to criterion 1a), namely whether the variance remains constant over the range of estimated mean values are shown in Figure 2.4a. Figure 2.4b depicts the p-values regarding criterion 1b), that is if the variance remains constant across dose groups. The p-values in Figure 2.4c concerns the normality assumption considered in criterion 2). In all cases high p-values support the validity of the assumption of variance homogeneity or normality whereas low p-values may indicate a violation. The variances of the summary statistics are given in Table 2.1. The variances are comparable within each column.

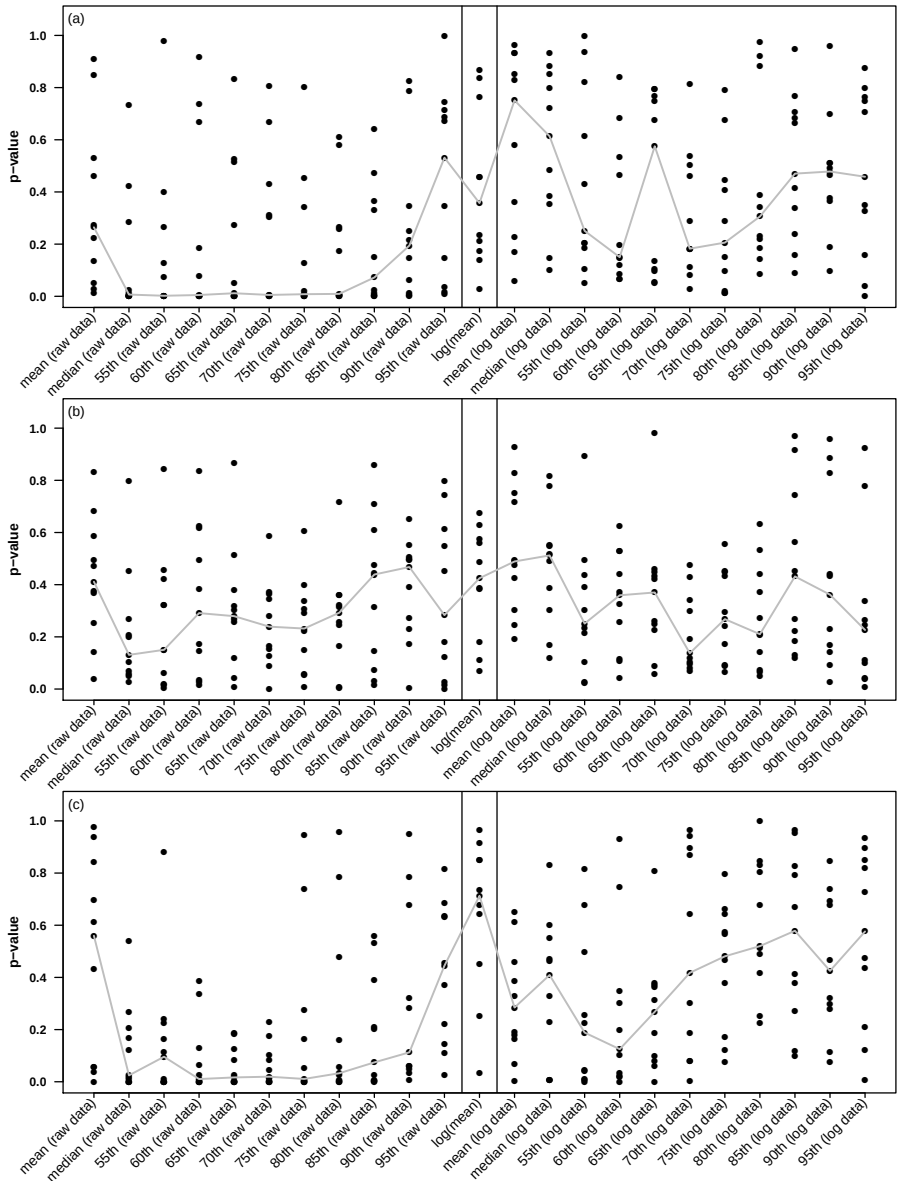

Figure 2.4: Assessment of the variance homogeneity and the normality assumption. The depicted p-values are concerned with (a) variance homogeneity over the range of estimated mean values, (b) variance homogeneity across dose groups and (c) normality. The median p-values are given by the grey lines and the two vertical lines separate the summary statistics of the raw data, the log(mean) and the summary statistics of the log-transformed data. The 'pth (raw data)' and 'pth (log data)' are short for the pth percentile of the raw data and of the log-transformed data, respectively.

Table 2.1: The variance of each summary statistic under the assumption that the within-sample distributions for each gel are log-normally distributed with $\mu = 2$, $\sigma^2 = 2$ and $n = 100$. The variances are multiplied by a factor of 50 for readability.

Summary statistic	Variance assuming a log-normal distribution (resembling raw data)	Variance assuming a normal distribution (resembling log(data))
mean	64.2	1.0
median	4.3	1.6
55th perc.	6.1	1.6
60th perc.	9.0	1.6
65th perc.	13.4	1.7
70th perc.	20.8	1.7
75th perc.	34.0	1.9
80th perc.	60.0	2.0
85th perc.	119.6	2.3
90th perc.	298.0	2.9
95th perc.	1272.5	4.7

In general, it is desirable to settle on a summary statistic that is associated with high p-values in Figure 2.4 and a low variance in Table 2.1. Furthermore, summary statistics having a meaningful interpretation are endeavored. The results in Figure 2.4 and Table 2.1 show that the summary statistic that most consistently accommodates these concerns is the median of the log-transformed data. Secondarily, the mean of the raw data seems to provide a suitable alternative. The remaining summary statistics either fall through in terms of the extracted p-values, the variance of the summary statistics or with respect to the ease of interpretation. Altogether, we conclude that median of the log-transformed data most suitably summarizes the % tail DNA distribution obtained from each slide.

2.4 Interpretation of the estimates when data are log-transformed

Whenever log-normally distributed data are encountered it can be advantageous to log-transform data in order to meet the assumptions of a range of standard analysis methods. However, this approach entails the challenge that the estimates from a statistical analysis are at a different scale than the original values, and this changes the interpretation. As the logarithmic transformation

is monotonic, the order of the data is preserved after transformation, hence

$$\text{median}(\log(\text{data})) = \log(\text{median}(\text{data})) \tag{2.8}$$

and the median of the log-transformed data can thus be viewed as a log-transformation of the median values. In the following we will outline how the estimates, as they often are obtained when a linear model is fitted in a statistical software package, can be interpreted on the original log-normal scale.

2.4.1 Moments of the log-normal distribution

A moment generating function (m.g.f.) is not defined for the log-normal distribution. Yet, the moments of the log-normal distribution can be obtained by means of the m.g.f. of the normal distribution.

Let $\boldsymbol{U} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{Y} = e^{\boldsymbol{U}}$ so that $\boldsymbol{Y} \sim \ln\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The joint moment of $\boldsymbol{Y}$ is

$$\mu'_{\boldsymbol{r}}(\boldsymbol{Y}) = \mathrm{E}(Y_1^{r_1} \cdots Y_k^{r_k}).$$

Substituting Y with e^U gives

$$\mu'_{\boldsymbol{r}}(\boldsymbol{Y}) = \mathrm{E}(e^{r_1 U_1} e^{r_2 U_2} \cdots e^{r_k U_k}) = \mathrm{E}(e^{\boldsymbol{r}^\top \boldsymbol{U}}),$$

however this is also the exact definition of the m.g.f. Since $\boldsymbol{U}$ is a normal random variable then

$$\text{m.g.f.} = e^{\boldsymbol{r}^\top \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{r}^\top \boldsymbol{\Sigma} \boldsymbol{r}}$$

and thus

$$\mu'_{\boldsymbol{r}}(\boldsymbol{Y}) = e^{\boldsymbol{r}^\top \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{r}^\top \boldsymbol{\Sigma} \boldsymbol{r}}. \tag{2.9}$$

(Kotz et al., 2000). From (2.9) the expected value (the first raw moment) is found as

$$\mathrm{E}(Y_i) = e^{\mu_i + \frac{1}{2}\Sigma_{ii}}, \tag{2.10}$$

and the covariance (the second bivariate mixed central moment) as

$$\begin{aligned} \mathrm{Cov}(Y_i, Y_j) &= e^{\mu_i + \mu_j + \frac{1}{2}(\Sigma_{ii} + \Sigma_{jj})}(e^{\Sigma_{ij}} - 1) \\ &= \mathrm{E}(Y_i)\mathrm{E}(Y_j)(e^{\Sigma_{ij}} - 1), \end{aligned} \tag{2.11}$$

which for $i = j$ reduces to the variance (the second central moment)

$$\mathrm{Var}(Y_i) = e^{2\mu_i + \Sigma_{ii}}(e^{\Sigma_{ii}} - 1)$$

From (2.11) we can obtain the elements of $\boldsymbol{\Sigma}$ as

$$\Sigma_{ij} = \ln\left(1 + \frac{\mathrm{Cov}(Y_i, Y_j)}{\mathrm{E}(Y_i)\mathrm{E}(Y_j)}\right),$$

whereas (2.10) provides

$$\begin{aligned}\mu_i &= \ln(\mathrm{E}(Y_i)) - \frac{1}{2}\Sigma_{ii} \\ &= \ln(\mathrm{E}(Y_i)) - \frac{1}{2}\ln\left(1 + \frac{\mathrm{Var}(Y_i)}{\mathrm{E}(Y_i)^2}\right)\end{aligned}$$

2.4.2 Relation between expected values of normal and log-normal random variables

Using the results about the moments of the log-normal distribution we can now establish the relation between estimates as they often are obtained when a linear model is fitted and the expected values on the original log-normal scale. Consider the ratio $\mathrm{E}(Y_i)/\mathrm{E}(Y_j)$. From (2.10) then

$$\begin{aligned}\frac{\mathrm{E}(Y_i)}{\mathrm{E}(Y_j)} &= \frac{e^{\mu_i + \frac{1}{2}\Sigma_{ii}}}{e^{\mu_j + \frac{1}{2}\Sigma_{jj}}} \\ &= e^{\mu_i + \frac{1}{2}\Sigma_{ii} - \mu_j - \frac{1}{2}\Sigma_{jj}}\end{aligned}$$

Taking the natural log of the ratio gives

$$\ln\left(\frac{\mathrm{E}(Y_i)}{\mathrm{E}(Y_j)}\right) = \mu_i + \frac{1}{2}\Sigma_{ii} - \mu_j - \frac{1}{2}\Sigma_{jj}$$

so that

$$\mu_i - \mu_j = \ln\left(\frac{\mathrm{E}(Y_i)}{\mathrm{E}(Y_j)}\right) - \frac{1}{2}\Sigma_{ii} + \frac{1}{2}\Sigma_{jj} \tag{2.12}$$

Often the variances Σ_{ii} and Σ_{jj} are assumed to be equal and (2.12) is reduced to

$$\mu_i - \mu_j = \ln\left(\frac{\mathrm{E}(Y_i)}{\mathrm{E}(Y_j)}\right) \tag{2.13}$$

In many statistical software packages, including R and SAS, the estimates are by default parameterized as the difference between the average responses for each group compared to a reference group. The result in (2.13) is thus relevant whenever data are log-transformed prior to the analysis. It means that if we for instance obtain an estimate that is $\mu_i - \mu_j = 0.69$, then on the original

scale (log-normally distributed data prior to the transformation) the ratio of the means is

$$\frac{\mathrm{E}(Y_i)}{\mathrm{E}(Y_j)} = e^{0.69} = 2,$$

that is, the response in group i is twice the size of the response in group j. It is important to keep in mind that the result given in (2.13) only is valid when the variances Σ_{ii} and Σ_{jj} are assumed to be equal. This assumption may not hold when e.g. a given compound is known to affect the variance, thereby inducing different variances in the groups receiving the compound compared to the vehicle group.

The result in (2.13) is quite convenient for at least two reasons. First, although the interpretation of the estimates deviates from the usual interpretation, it is still relatively straightforward to interpret the results. Second, since the estimates are translated to a ratio on the original scale it means that the estimates are invariant to the reference level, which for Comet assay studies is the vehicle group. It can be quite convenient and the ratio of expected values, phrased as *fold change*, is sometimes used to communicate the results from Comet assay studies (see e.g. Smith et al., 2008; Lovell and Omori, 2008; Guérard et al., 2014). The invariance can be seen as a fortunate property for log-normally distributed data that does not apply to normally distributed data.

Chapter 3

Type I errors

Statistical tests often concern whether a specified hypothesis is to be rejected or not. However, when inference is made on the grounds of observed data its validity cannot be absolutely guaranteed. Unavoidably, there is a risk of committing an error when conclusions are drawn and mainly two types of errors are accentuated. A type I error is a false rejection of the hypothesis in question, whereas a type II error occurs when we incorrectly fail to reject the hypothesis. Many statistical tests involve that a bearable type I error rate is specified, which also is known as the nominal α. The current chapter deals with this type I error rate, while Chapter 4 concerns the type II error rate both in a general setting and within the framework of Comet assay studies.

In practice, the actual type I error rate can deviate from the nominal α, which is a concern that recurrently has been addressed in the literature. A frequent source is the lack of correction when multiple hypotheses are tested, which also is the main concern in the perpetration of data dredging (Strasak et al., 2007; Nuzzo, 2014). When model assumptions are violated it likewise affect the type I error rate, and the severity depends on the type of assumption and the degree to which it is violated.

As described in Section 2.1 data from *in vivo* Comet assay studies usually are hierarchically structured. During our work with *in vivo* Comet assay data we realized that this hierarchical structure does not seem to be consistently reflected

in the statistical models as they are reported in the Comet assay literature. Consequently the assumption regarding independence among observations is violated. This assumption is rather crucial and the implications when it is violated are severe.

This chapter addresses the implications when the hierarchical structure is ignored. This issue has been addressed previously (e.g. Kenny and Judd, 1986; Kromrey and Dickinson, 1996; Baldwin et al., 2005; Musca et al., 2011) but appears to continue to prevail. Our aim with this study is to participate in raising awareness on this serious matter and to examine in which way the implications arise. Paper A and B deal with the implications specifically in the framework of Comet assay studies. The hierarchical design considered here is by no means limited to Comet assay studies though and commonly occurs in a variety of fields. Accordingly Paper C presents this issue in a general framework.

3.1 Statistical analysis of Comet assay studies: A literature study

To assess to which extent the hierarchical structure of Comet assay studies is disregarded a literature study was conducted. More details are provided in Paper B.

Papers were retrieved from the search engine Web of Science with *title: in vivo* and *topic: Comet assay* from January 2012 until December 2013, which resulted in 95 papers. Of these, 47 papers conducted *in vivo* Comet assay studies with an experimental setup as in Figure 2.3 and accordingly were included in the literature study.

It was in general not easy to determine how the statistical analysis was conducted as the description of the statistical methodology often was brief and imprecise. Most papers used a one-way ANOVA (45%), ANOVA (21%) or Kruskal-Wallis test (15%), i.e. methods that assume independence. 18 papers (38%) stated "Results are expressed as mean $\pm$ SD" (or mean $\pm$ SE). However, it was not clear how it was calculated (i.e. for each slide, for each animal etc.) or if the statement applies to the data representation in tables or in the statistical analysis. 23 papers (49%) calculated a summary statistic prior to the statistical analysis, but of these it was only clear how it was done in 15 papers (65%). Overall, in 24 papers (51%) it seemed as no summary statistic was calculated prior to the statistical analysis. The 24 papers were published in 20 different biomedical journals.

We find this result problematic for two reasons. First, the brief and imprecise description of the statistical methodology impedes reproducibility as well as a proper interpretation of the results. Second, the lack of a calculated summary statistic combined with the reported statistical methods strongly indicate that the hierarchical structure in some cases is not suitably reflected in the statistical analysis.

3.2 The type I error rate disregarding the hierarchical structure of data

The following section addresses the implications when the hierarchical structure of data is disregarded. First, approximate closed-form expressions for the type I error rate are derived. Next, the validity of the approximate type I error rates are validated by a simulation study. A more elaborate exposition of all derivations and results are provided in Paper B and C.

3.2.1 Sampling distributions

When hierarchical data structured as in Figure 2.3 is fitted a one-way ANOVA as defined in (2.4) an F-statistic is calculated as

$$F_{\text{anova}} = \frac{bn \sum_{i=1}^{a} \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot}\right)^2 / (a-1)}{\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} \left(Y_{ijk} - \overline{Y}_{i\cdot\cdot}\right)^2 / (a(bn-1))} \tag{3.1}$$

The sum of squares in the numerator is distributed as

$$bn \sum_{i=1}^{a} \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot}\right)^2 \sim \left(n\sigma_\beta^2 + \sigma^2\right) \chi^2 \left(a-1, \lambda\right),$$

where

$$\lambda = \frac{bn \sum_{i=1}^{a} \tau_i^2}{n\sigma_\beta^2 + \sigma^2}.$$

Since Y_{ijk} are not independent the sum of squares in the denominator of (3.1) does not follow a usual χ^2-distribution (see Box (1954) for details) but is rather a linear combination of independent χ^2-distributed random variables. An approximate distribution can be obtained by means of the Welch-Satterthwaite

approximation (Welch, 1938; Satterthwaite, 1941; Box, 1954) in that the sum of squares are approximated by a scaled χ^2-distribution

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk}-\overline{Y}_{i\cdot\cdot}\right)^2 \dot{\sim} c\chi^2(\nu)$$

Matching the first two moments leads to

$$c=\frac{(b-1)(n\sigma_\beta^2+\sigma^2)^2+b(n-1)(\sigma^2)^2}{(b-1)(n\sigma_\beta^2+\sigma^2)+b(n-1)\sigma^2}$$

and

$$\nu=\frac{a((b-1)(n\sigma_\beta^2+\sigma^2)+b(n-1)\sigma^2)^2}{(b-1)(n\sigma_\beta^2+\sigma^2)^2+b(n-1)(\sigma^2)^2},$$

where ν is known as the effective degrees of freedom (Satterthwaite, 1941). The ratio of χ^2-distributed random variables each divided by their degrees of freedom follows an F-distribution. However, to assess the sampling distribution of the test statistic in (3.1) we need to adjust for the incorrect degrees of freedom in the denominator and also that the numerator and denominator are scaled differently. This leads to

$$F_{\text{anova}} \dot{\sim} \xi\, F(a-1,\nu,\lambda), \tag{3.2}$$

where

$$\xi=\frac{(bn-1)(n\sigma_\beta^2+\sigma^2)}{(b-1)(n\sigma_\beta^2+\sigma^2)+b(n-1)\sigma^2}.$$

Accordingly, the expected value becomes

$$E(F_{\text{anova}}) \approx \xi\,\frac{\nu}{(a-1)(\nu-2)}\left(a-1+\frac{bn\sum_{i=1}^{a}\tau_i^2}{n\sigma_\beta^2+\sigma^2}\right) \tag{3.3}$$

(Johnson et al., 1995). When $\sigma_\beta^2=0$ then $\xi=1$ and under H_0 and for sufficiently large ν then $E(F_{\text{anova}})\approx 1$. For $\sigma_\beta^2>0$ then $\xi>1$ implying that $E(F_{\text{anova}})>1$.

The observed F_{anova}-statistic is improperly compared to a critical value from an unscaled F-distribution and the type I error rate is thus found as

$$\text{Type I error rate} \approx 1-G_s(F_{\alpha;a-1,a(bn-1)};a-1,\nu),$$

where G_s refers to the scaled cumulative distribution function of F_{anova}. Equivalently, the type I error rate can be found based on a non-scaled cumulative distribution

$$\text{Type I error rate} \approx 1-G\left(\xi^{-1}\,F_{\alpha;a-1,a(bn-1)};a-1,\nu\right) \tag{3.4}$$

and the type I error rate can be obtained by means of a non-scaled F-distribution, which is readily available in most statistical software.

The type I error rate can also be formulated in terms of the ratio of the animal and error variance component

$$\sigma^2_{\text{ratio}} = \frac{\sigma^2_\beta}{\sigma^2}$$

The effective degrees of freedom and the inverse scaling factor then become

$$\nu = \frac{a((b-1)(n\sigma^2_{\text{ratio}}+1)+b(n-1))^2}{(b-1)(n\sigma^2_{\text{ratio}}+1)^2+b(n-1)}$$

and

$$\xi = \frac{(bn-1)(n\sigma^2_{\text{ratio}}+1)}{(b-1)(n\sigma^2_{\text{ratio}}+1)+b(n-1)}.$$

3.2.2 Simulation study

A simulation study was carried out to verify the results from the closed-form expressions given in the previous section. Also, the impact on the type I error rate when the hierarchical structure is disregarded is provided for different sample sizes and variance ratios.

The particular hierarchical data structure considered her occurs within a wide range of scientific disciplines. The levels of the parameters are selected to comprise the diversity as they naturally appear in these different fields. This section recapitulates some of the results of Paper C, while type I error rates reflecting parameter values specifically relevant for Comet assay studies are given in Paper A and B. Due to the more general setting of Paper C the variables *animals* and *slides* are below referred to as *groups* and *observations*, respectively.

Table 3.1 outlines the type I error rates for different combinations of treatments, groups per treatment, observations per group and ratios of the variance components. The approximate type I error rates are calculated from (3.4), whereas the simulated type I error rates are obtained by simulating data from model (2.2) with 10000 simulations conducted for each combination. The type I error rate in general increases with increasing number of treatments, number of observations per group and the variance ratio, $\sigma^2_{\text{ratio}} = \sigma^2_\beta/\sigma^2$. Surprisingly, the number of groups per treatment do not noticeably affect the type I error. For all cases the type I error rate was greater than the nominal α at 0.05. Most combinations gave type I error rates greater than 0.10 and more than half resulted in

Table 3.1: Type I error rates for different combinations of number of treatments (a), groups per treatment (b), observations per group (n) and variance ratio $\sigma^2_{\text{ratio}} = \sigma^2_\beta/\sigma^2$. The approximate type I error rates were found from (3.2), and the simulated type I error rates were based on 10000 simulations for each combination (each cell). In all cases the nominal α was 0.05.

Treatm.	Groups	Observ.	σ^2_{ratio}							
			0.25		0.50		1.00		2.00	
			Appr.	Sim.	Appr.	Sim.	Appr.	Sim.	Appr.	Sim.
2	2	2	0.076	0.074	0.099	0.095	0.134	0.127	0.177*	0.161
		10	**0.259**	**0.255**	**0.363**	**0.366**	**0.465**	**0.462**	**0.548**	**0.547**
		100	**0.683**	**0.682**	**0.760**	**0.765**	**0.813**	**0.816**	**0.849**	**0.847**
	50	2	0.074	0.074	0.090	0.087	0.110	0.108	0.130	0.131
		10	**0.242**	**0.240**	**0.328**	**0.332**	**0.406**	**0.405**	**0.462**	**0.457**
		100	**0.668**	**0.663**	**0.738**	**0.737**	**0.784**	**0.783**	**0.812**	**0.813**
6	2	2	0.107	0.107	0.160	0.161	**0.242**	**0.235**	**0.335**	**0.336**
		10	**0.594**	**0.590**	**0.788**	**0.794**	**0.899**	**0.901**	**0.948**	**0.948**
		100	**0.993**	**0.992**	**0.998**	**0.998**	**0.999**	**1.000**	**1.000**	**1.000**
	50	2	0.101	0.102	0.141	0.140	0.195	0.196	**0.251**	**0.249**
		10	**0.558**	**0.565**	**0.738**	**0.734**	**0.849**	**0.851**	**0.905**	**0.901**
		100	**0.991**	**0.990**	**0.997**	**0.998**	**0.999**	**0.999**	**0.999**	**1.000**

* Approximate type I error rates not covered by the 95% confidence intervals of the simulated type I error rates.
Type I error rates greater than 0.20 are marked in bold.

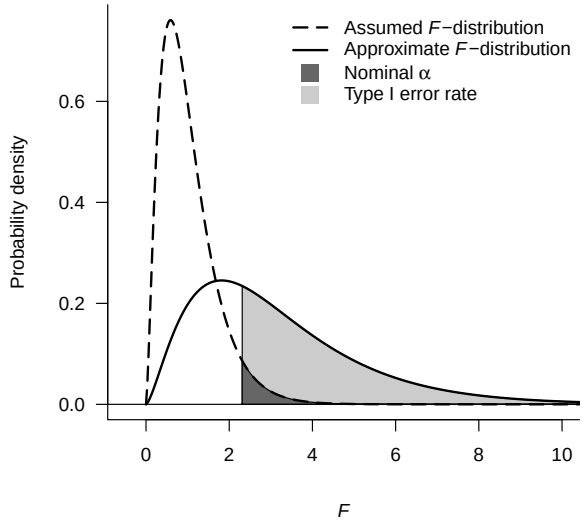

Figure 3.1: The F-distributions in case of six treatments, two groups per treatment, ten observations per group and $\sigma^2_{\text{ratio}} = 0.25$. The assumed F-distribution refers to the distribution from which the critical value is obtained. The approximate F-distribution is the distribution of F_{anova} as defined in (3.2). The approximate F-distribution has a heavier right tail implying that the type I error rate is greater than the nominal α at 0.05.

type I error rates greater than 0.50. To assess the validity of the approximate type I error rates they were informally compared to the Wilson 95% confidence intervals (CI) (Wilson, 1927; Agresti and Coull, 1998) of the simulated type I error rates. In all but one case the approximate type I error rates were covered by the CI for the simulated type I error rates. This is what is expected given the number of comparisons and the confidence level. The one case not covered by the appertaining CI is marked with an asterisk in Table 3.1.

Figure 3.1 illustrates the inflation of the type I error rate. Here the sampling distributions are shown for six treatments, two groups per treatment, ten observations per group and a variance ratio of $\sigma^2_{\text{ratio}} = 0.25$. The dashed line is the assumed sampling distribution of F_{anova} when the hierarchical structure is ignored, and the distribution from which the critical value is determined. The solid line outlines the approximate scaled F-distribution given in (3.2). It is seen that additional skewness is imposed on the approximate F-distribution implying a heavier right tail. For the combinations shown in Table 3.1 the expectations $E(F_{\text{anova}})$ given in (3.3) are between 1.21 and 134.56.

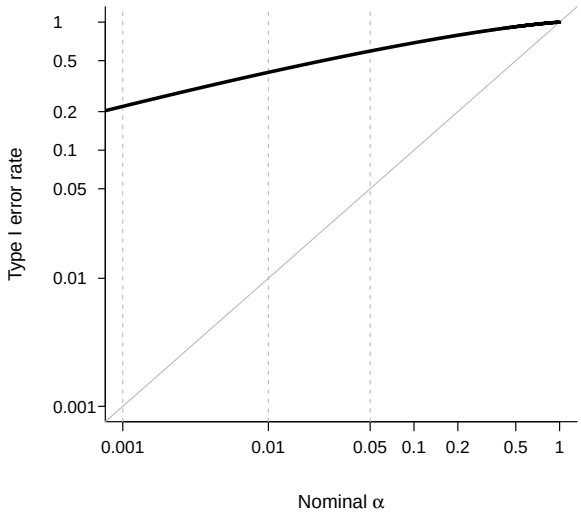

Figure 3.2: The type I error rate versus the nominal α (black line) in case of six treatments, two groups per treatment, ten observations per group and $\sigma^2_{\text{ratio}} = 0.25$. The grey solid reference line corresponds to equality between the type I error rate and the nominal α. Both axes are on a logarithmic scale.

Proceeding with the same example Figure 3.2 shows the type I error rate as a function of the nominal α (black line) on the logarithmic scale. The grey solid line outlines equality between the type I error rate and the nominal α. For all values of the nominal α the type I error rate is seen to be considerably inflated when the hierarchical structure is ignored. For the example considered here a nominal α of 0.001 corresponds to a type I error rate of 0.220 while a nominal α of 0.01 corresponds to a type I error rate of 0.406. That means that even when significance is demonstrated at a level of 0.001, which often is considered fairly strong evidence against H_0, it does not guarantee that the actual type I error rate is anywhere near the conventional level at 0.05 if a hierarchical structure has been disregarded.

Chapter 4

Type II errors

The previous chapter concerned type I errors, which take place when the null hypothesis incorrectly is rejected. There is another type of error that is closely related to the type I error, namely the type II error that occurs when we fail to reject a null hypothesis when it in fact does not hold.

Committing type I error is often considered more serious than committing a type II error. Statistical tests in general control the type I error rate by means of the significance level, α, which explicitly must be specified by the data analyst. In contrast, it is possible to conduct a statistical test without considering the type II error rate by any means. This may partly explain the lack of consideration toward controlling the type II error rate that characterize too many biomedical studies (e.g. Carp, 2012; Button et al., 2013; Gaskin and Happell, 2013; Koletsi et al., 2014).

This chapter will be dealing with type II error rates and related quantities in general and for *in vivo* Comet assay studies in particular. *In vivo* Comet assay studies are conducted using living animals, usually mice or rats, and accordingly they belong to a significant area of biomedical studies, namely animal experiments. The diverse and far-reaching impact of animal experiments is emphasized by Hau and Schapiro (2014, p. vii):

> "Most of our present knowledge concerning human physiology, mi-

> crobiology, immunology, pharmacology, pathology, and related disciplines has been gained from studies involving animals [...] Biomedical research involving animals remain absolutely essential for the advancement of the medical, veterinary, agricultural and biological sciences. All drugs prescribed for use in humans and animals have been developed and tested in laboratory animals as models. Noninvasive imaging techniques are optimized in animal models. New surgical techniques and materials are evaluated in animals before they are applied in cases that involve humans or domestic animals. The dramatic developments in genetics – the sequencing of the human genome and the genomes of many of the most important laboratory animal species, translational research, and personalized medicines – all rely on access to high-quality laboratory animals as models for humans."

Nevertheless, animal testing is a matter of quite some controversy. While some consider it an indispensable tool toward biomedical advances other see it as plain unjustifiable animal cruelty. Emotions and passion are found on both sides of this ongoing dispute. The discrepancies are not easily accommodated and existing legislation reflect a compromise that strives to comply with both sides. One principle that has become widely accepted is the The Three Rs (3Rs) (Russell and Burch, 1959; Hau and Schapiro, 2014), which is short for Replacement, Reduction and Refinement. This means that animal testing should only be carried out when no other options are available (Replacement), the number of animals used should be reduced as much as possible (Reduction) and the experiments should be refined in order to reduce the discomfort of the animals (Refinement). The 3R principle has since 2010 been covered by the EU legislation (European Parliament, 2010) and a 3R-Center has been appointed by the Ministry of Food, Agriculture and Fisheries of Denmark[1].

One factor that affects the type II error rate is the sample size as will be outlined in Section 4.2. This is addressed by the second R, that is, Reduction. The sample size in many cases directly impacts the resources that need to be allocated a given study; hence the sample size is not uncommonly a concern in many research areas. However, it is a rare exception that such an experimental design issue is of a concern to as different parties as private individuals, organizations and supreme authorities. The attention is a natural reflection of the major ethical dilemmas that animal experiments are entangled in. These very ethical concerns are the reason that the sample size must be pondered upon carefully.

[1] see e.g. http://fvm.dk/landbrug/indsatsomraader/dyrevelfaerd-og-transport/3r/3r-center/ (in danish) and http://www.foedevarestyrelsen.dk/english/Animal/Pages/The-Danish-3R-Center.aspx (in english)

It is possible that the immense focus on reducing the number of animals have been implemented a little too well within some biomedical research areas. Although it has not led to that statistical power in general are determined prior to the animal studies (Hawkins et al., 2013), it instead seems that numerous animal experiments make use of small sample sizes without providing any explicit justification, leading to that statistical power for these studies are insufferably low (see e.g. Jennions and Møller, 2003; Hofmeister et al., 2007; Sena et al., 2010; Smith et al., 2011; Giuffrida, 2014). While an excessive use of animals clearly is a waste of animal lives, it may not be so obvious that the opposite can be of even greater waste. The type II error rates are affected by the sample size and such incidences erroneously indicate[2] no effect of the treatment in question. Studies relying on small sample sizes may thus not only be uninformative but even prove misleading.

Due to the ongoing emotional dispute it may not be easy to encourage compliance toward the use of larger sample sizes when reasonable. Statistical power provides a sensible reasoning for choosing a proper sample size, be it greater or smaller, rather than using what sometimes appears to be an arbitrary number of animals. Given the importance of animal studies, the ethical concerns and the massive resources allocated these, we find it paramount to draw attention to power and sample size determinations in general and for animal studies in particular.

4.1 Type II errors and statistical power

The concept of statistical power dates back to the late 1920s. Statistical power is closely related to the type II error rate (β), in that

$$\text{Power} = 1 - \beta, \tag{4.1}$$

that is, statistical power is the probability of correctly rejecting the null hypothesis. It naturally emerges within the framework of hypothesis testing as it was formulated by Jerzy Neyman and Egon Pearson (Neyman and Pearson, 1928a,b). The hypothesis tests (or the related concept of significance tests in words of Fisher) as they often are presented nowadays is an ungainly mixture of ideas presented by Ronald Fisher on one side and Neyman and Pearson on the other side (Hubbard and Bayarri, 2003; Nuzzo, 2014). While Fisher refused to contemplate any hypotheses but the null, Neyman and Pearson suggested the explicit definition of two competing hypotheses; the null hypothesis and the

[2] although the null hypothesis has not been proved when we fail to reject it, in practice it is sometimes perceived as such (Altman and Bland, 1995).

alternative. Type I error rates can be found solely on grounds of the null hypothesis, whereas type II error rates and statistical power requires the explicit formulation of an alternative hypothesis. The idea of statistical power did not immediately gain currency. Its revival in the early 1960s is mainly attributable to the significant work by Jacob Cohen (Descôteaux, 2007). Cohen published several papers and books on this topic, of which several became highly cited classics especially in the field of behavioral science (Cohen, 1988, 1992, 2005). Although Cohen succeeded in raising awareness on statistical power, his ideas on standardized small, medium and large effect sizes have been subjected to criticism (Lenth, 2001).

Nowadays, a brief introduction to the concept of statistical power forms part of the curriculum in many introductory statistics courses and power calculations are in general recommended whenever a study is to be conducted. However, as mentioned above disparity seems to exist between this recommendation and common practice within the biomedical field. This inconsistency may result from several factors. First, it is possible that researchers are inspired by the work of others. If relevant literature does not report the use of power calculations, they may less likely be used prospectively. Second, assumptions about the effect size of interest and variability of data must be made as part of the power determinations, i.e. before the study is carried out, which inherently is difficult. Third, power calculations add considerable to the statistical complexity. Whereas hypothesis testing only requires the comparison of a test statistic to a null distribution, power determinations further include the formulation of an alternative distribution, which almost always imply non-central distributions. Fourth, the literature on statistical power is sparingly scattered, making this topic easy to neglect.

Owing to the wide range of statistical software available it is possible to apply statistical methods without understanding the underlying theory in depth. Numerous options are also available for assessing statistical power. Functions for determining statistical power are implemented in all major software packages such as SAS, R, SPSS, Minitab, Stata etc. Unfortunately, some packages provide the power software in add-on packages that must be purchased separately and is thus not available to all researchers. Also, the terminology is not always consistent and can be difficult to decode. Our own experience is primarily with R that provides power calculations in several different packages (e.g. Halvorsen, 2013; Fan, 2013; Anderson, 2014; Rizopoulos and Tsonaka, 2009; Millard, 2013; Ball, 2012; Labes, 2014). Some of them overlap in the power functions they provide and the input arguments happen to differ from package to package, e.g. power for a *t*-test using `power.t.test` from the `stats` package (R Core Team, 2014) requires the specification of an *unstandardized* effect size (difference in means), while the equivalent `pwr.t.test` from the `pwr` package (Champely, 2012) expects the *standardized* effect size (which is the unstandardised effect size di-

vided by the standard deviation). Such subtle though crucial differences that are ubiquitous in the literature, available online power calculators and commercial software add to the confusion over statistical power. It is beyond doubt that power software developers are seeking to promote the feasibility of power calculations and that it merely seems impossible not to contribute to the confusion due to the lack of a consistent terminology that permeates this framework. Two software packages that stand out though are *Java Applets for Power and Sample Size*[3] (Lenth, 2009) and *G*Power*[4] (Buchner et al., 2014). The authors have developed software that is versatile and are supplied with manuals and papers (Lenth, 2001, 2007; Faul et al., 2007; Mayr et al., 2007; Faul et al., 2009; The G*Power Team, 2014). This combination enables a consistent terminology within the material provided by each developer group. Statistical power can also be assessed by means of simulation, which is an obvious choice for many statisticians and data analysts due to the simplicity and versatility. However, this approach requires knowledge on the statistical model in question and about stochastic simulation in general. For biomedical researchers to pursue this path there still is a need for an introduction to the general idea underlying statistical power. An excellent introduction on this matter is provided by Bolker (2008).

Considering the importance of power calculations there is not much literature covering this area compared to the wealth of expositions that continuously are published on a range of hypothesis tests. Some papers address this topic by reviewing how power calculations are used (or rather not used) in publications (e.g. Hofmeister et al., 2007; Smith et al., 2011; Giuffrida, 2014) thereby raising awareness on this topic. Others provide closed-form expressions or power curves for standard tests, such as the t-test or a one-way ANOVA (e.g. Chow et al., 2002; Algina and Olejnik, 2003; Livingston and Cassidy, 2005); among these Julious (2004) provides a remarkably comprehensive tutorial. While they all genuinely contribute to an improvement concerning the use of power calculations in biomedical studies, we still find it inherently difficult to learn the basic concept of statistical power based on the current literature. We have not yet seen any papers or textbooks expositing the theory of statistical power by illustrating the null and the alternative distribution comprising an introduction to non-central distributions. We consider this a natural path of dissemination as statistical power is directly derived from these two distributions. As part of this PhD project a tutorial paper on statistical power was therefore initiated. As it is not yet ready for submission it is not attached in the Appendix, but selected parts are included in the following section. The intended audience is researchers, including biomedical researchers, that not are statisticians but who have a well-founded basic knowledge in statistics, i.e. a thorough understanding of sampling distributions and p-values.

[3] available at http://homepage.stat.uiowa.edu/ rlenth/Power/ (online or stand-alone)
[4] available at http://www.gpower.hhu.de/ (stand-alone)

4.2 Summary of a tutorial paper on statistical power

4.2.1 Introduction

When performing a hypothesis test, we either reject the null hypothesis or fail to reject it. The decision is based on the significance level, α, which is fixed prior to the conduction of the hypothesis test. There are four possible outcomes when conducting the hypothesis test: 1) correctly rejecting the null hypothesis, 2) correctly failing to reject the null hypothesis, 3) incorrectly rejecting the null hypothesis (type I error) or 4) incorrectly failing to reject the null hypothesis (type II error). The probability of committing a type I error is called α, whereas the probability of committing a type II error is β. Statistical power is closely related to β in that

$$\text{power} = 1 - \beta$$

Statistical power is thus the probability of correctly rejecting the null hypothesis (Johnson et al., 2010).

	reject H_0	fail to reject H_0
H_0 is true	type I error (α)	correct failure of rejection
H_0 is false	correct rejection (power)	type II error (β)

Table 4.1: The four possible outcomes of a hypothesis test. The probability of the outcomes are seen in parentheses

The type I error, α, is familiar to most people applying statistical hypothesis tests, since this figure often serves as criteria for the rejection or failure of rejection of the null hypothesis. The type II error, β, is also important to control, since incidences of these erroneously may be perceived to indicate no effect of the treatment in question. Poorly designed studies with a high rate of type II errors may thus not only be uninformative but even prove misleading. Nevertheless, considerations toward this issue often seems to be disregarded indicating a low or absent awareness of the possibility of committing this type of error. Accordingly, power seems to be equally unappreciated and too often not considered in the experimental design phase (Button et al., 2013).

4.2.2 The concept of power

In order to comprehend the concept of power, it may be useful to start with the simplest possible case. This section will consider the hypothesis test of a single sample mean in the normal case with known variance. Throughout this exposition H_A refers to a general alternative hypothesis, and $H_i, i \in \mathbb{N}_{>0}$ refer to specific alternative hypotheses.

Suppose a random sample, $X_i, i = 1, ..., n$ is drawn from a normal population with mean μ and known variance σ^2. The sample mean $\overline{X}$ is then assumed to be a random variable following a normal distribution with mean μ and variance σ^2/n.

We now wish to test the null hypothesis $H_0 : \mu = \mu_0$ versus the alternative hypothesis H_A: $\mu > \mu_0$. In Figure 4.1 the distribution of $\overline{X}$ under H_0 and a specific alternative distribution with $\mu = \mu_1$ is shown. The vertical black line indicates the critical value that is obtained from the null distribution. α is the area to the right of the critical value under H_0 (the probability of incorrectly rejecting H_0), β is the area to the left of the critical value under H_1 (the probability of incorrectly failing to reject the null hypothesis) and power is the area to the right of the critical under H_1 (the probability of correctly rejecting H_0).

This plot illustrates why a number of assumptions must be addressed in order to estimate **power**:

Distance: The distance between the true mean μ and the hypothesized mean μ_0 corresponds to the *unstandardized* effect size as described by Cohen (1988). Increasing the distance entails a better separation of the distributions of $\overline{X}$ under H_0 and H_1, respectively, thus increasing power as illustrated in the first row of Figure 4.1.

Variability: Since the variance of $\overline{X}$ is σ^2/n, the distributions of $\overline{X}$ under H_0 and H_1 become more narrow as the variability in data decreases, hence providing better power. Lowering the variability in data can be obtained through e.g. employment of a proper experimental design.

Sample size: The variance of $\overline{X}$ is diminished by increasing the sample size in turn increasing power, cf. previous item. The role of variability and sample size on power is seen in the middle row of Figure 4.1.

Level of significance: The critical value is determined based on the specified significance level (α), which in turn affects power. As seen in last row of Figure 4.1, lowering α decreases the power and vice versa.

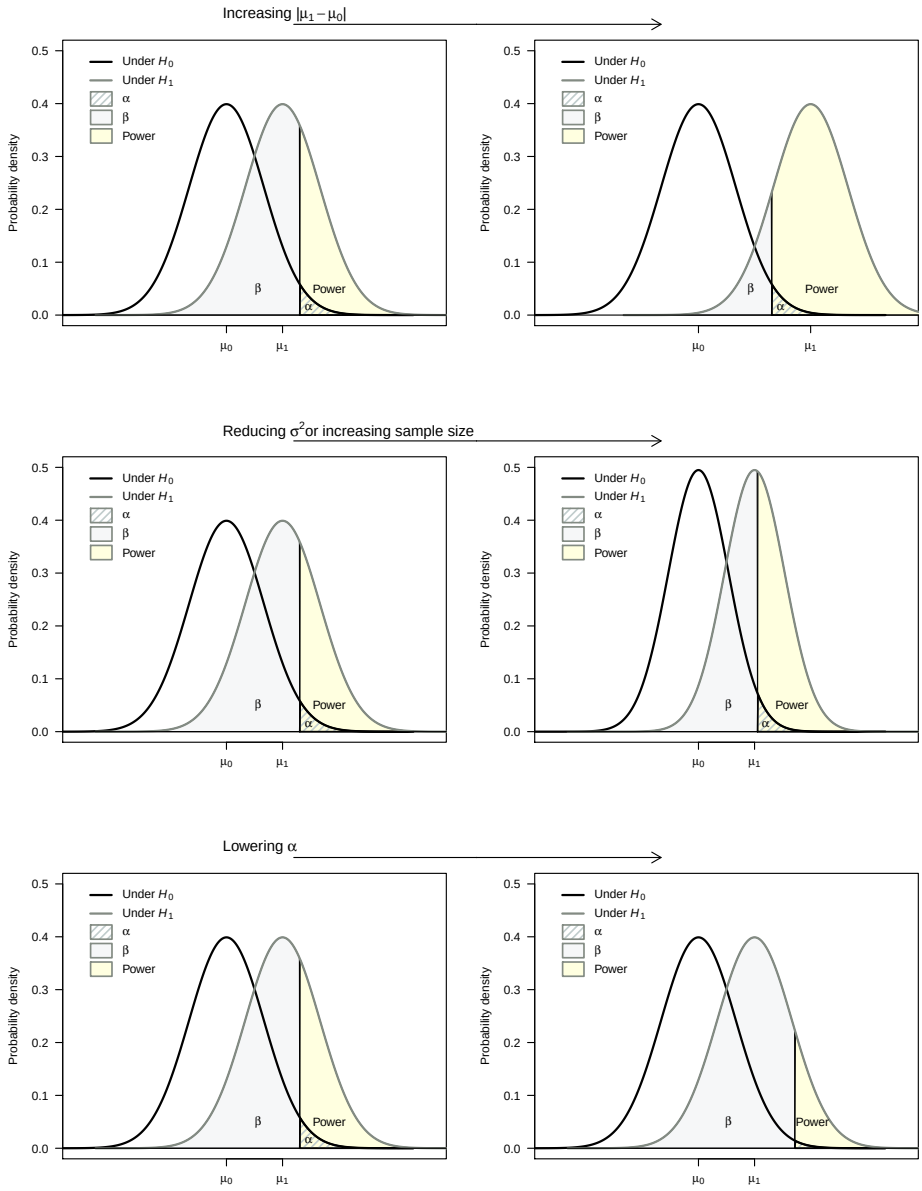

Figure 4.1: The sampling distributions under H_0 and H_1 with the areas of α, β and power outlined on the graphs. The black vertical line indicates the critical value. The upper row illustrates how a change in the distance $|\mu_1 - \mu_0|$ affects power. Here it is seen how an increased distance leads to increased power. The middle row shows how the width of the sampling distributions change when the variability or the sample size are altered. This example shows how decreased variability or an increased sample size lead to increased power. The last row shows how power changes when the significance level, α, is changed. Here it is shown how a decreased significance level leads to a reduction of power.

Others: In less simplified cases (many real-life cases) further assumptions must be made, e.g. about the correlation of repeated measurements etc.

The listed quantities in bold are related so that any of them can be determined based on the others. Perhaps the most common way of conducting power analysis is to provide reasonable estimates of requisite quantities, and then determine the sample size necessary to obtain the appropriate power.

Note that Figure 4.1 illustrates just one out of (infinitely many) alternative hypotheses and is valid only in the specific case where the true mean equals μ_1. Hence, a power analysis is conducted for one specific hypothesized distance between μ_0 and μ_1 (while holding the other factors listed above fixed).

Calculating power

To calculate power we determine the critical value(s) obtained from the distribution under H_0, and use it/them to obtain the critical region(s) of the alternative distribution (the number of critical values/regions correspond to the number of sides in the hypothesis test). This can be done directly using the distributions of the raw sample means, but in order to generalize the formulas the random variable $\overline{X}$ is standardized.

Under H_0 the standardized random variable

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \tag{4.2}$$

whereas under H_1

$$Z \sim N\left(\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}, 1\right) \tag{4.3}$$

The power for a one-sided test is then obtained as

$$\text{power} = 1 - \beta = 1 - \Phi\left(z_\alpha - \frac{|\mu_1 - \mu_0|}{\sigma/\sqrt{n}}\right) \tag{4.4}$$

Note that (4.4) always returns the area of the critical region corresponding to the direction of the hypothesis test thus covering both the alternatives *greater* and *less*.

For a two-sided hypothesis test

$$\text{power} = 1 - \Phi\left(z_{\alpha/2} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) + \Phi\left(-z_{\alpha/2} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)$$

Approximate power for a two-sided hypothesis test is obtained by ignoring a small value $\leqslant \alpha/2$

$$\text{power} \sim 1 - \Phi\left(z_{\alpha/2} - \frac{|\mu_1 - \mu_0|}{\sigma/\sqrt{n}}\right)$$

4.2.3 Non-central distributions

Non-central distributions arise in power and sample size calculations apart from a few exceptions. Non-central distributions are generalizations of the appertaining (central) distributions. An example is the non-central t-distribution, which is a generalization of the central t-distribution. Often the term 'central' is omitted and the central t-distribution is simply referred to as the t-distribution.

Non-central χ^2-, t- and F-distributions appear in power calculations for some of the most widely used models, such as the t-test and ANOVA.

Non-central χ^2-distribution

Assume that $U_1, U_2, ..., U_\nu$ are independent variables each following a standard normal distribution and $\delta_1, \delta_2, ..., \delta_\nu$ are constants. Then

$$V = \sum_{i=1}^{\nu} (U_i + \delta_i)^2$$

follows a non-central χ^2 distribution with ν degrees of freedom and non-centrality parameter $\lambda = \sum_{i=1}^{\nu} \delta_i^2$, denoted $\chi^2(\nu, \lambda)$. The central χ^2-distribution is a special case of the non-central χ^2-distribution where $\lambda = 0$ (Johnson et al., 1994, chapter 29).

Non-central t-distribution

Assume that Z and V are independent random variables following a normal distribution with mean λ and variance 1 and a central χ^2-distribution with ν degrees of freedom, respectively. Then

$$T = \frac{Z}{\sqrt{V/\nu}} \tag{4.5}$$

follows a non-central t-distribution with ν degrees of freedom and non-centrality parameter λ, denoted $t(\nu, \lambda)$. When $\lambda = 0$ the non-central t-distribution becomes the central t-distribution (Johnson et al., 1994, chapter 31).

Non-central F-distribution

Consider the independent random variables, V_1 and V_2, each following a non-central χ^2-distribution with ν_1 and ν_2 degrees of freedom and non-centrality parameters λ_1 and λ_2, respectively. Then

$$W = \frac{V_1/\nu_1}{V_2/\nu_2}$$

follows a doubly non-central F-distribution with ν_1, ν_2 degrees of freedom and non-centrality parameters λ_1 and λ_2, denoted $F(\nu_1, \nu_2, \lambda_1, \lambda_2)$.

If V_1 follows a non-central χ^2-distribution with non-centrality parameter λ and V_2 follows a central χ^2-distribution, then W follows a singly non-central F-distribution with ν_1, ν_2 degrees of freedom and non-centrality parameter λ, denoted $F(\nu_1, \nu_2, \lambda)$. Often, this distribution is merely referred to as a non-central F-distribution thus leaving out the term 'singly', and this terminology also applies here.

When both V_1 and V_2 follow a central χ^2-distribution, then W follows a central F-distribution (Johnson et al., 1994, chapter 30).

4.2.4 Power and sample size calculations

We will now turn to the actual power calculations for some standard models often encountered. The first subsection will continue the example from the previous section treating a single sample with known variance, whereas more complex models are introduced gradually in later subsections.

[...subsection left out...]

One-sample t-test (unknown variance)

Suppose a random sample, $X_i, i = 1, ..., n$ is drawn from a normal population with mean μ and unknown variance σ^2. The sample mean $\overline{X}$ is a random

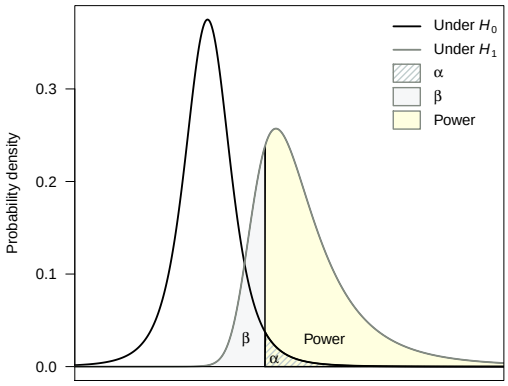

Figure 4.2: The test statistic, T, follows a *central* t-distribution under H_0 and a *non-central* t-distribution under H_1.

variable assumed to follow a normal distribution with mean μ and variance σ^2/n. In section 4.2.2 power for the test of a one-sample mean with known variance was considered. In many real-life applications the assumption of a known population variance is unrealistic though, and it must be estimated from prior studies, e.g. a pilot study.

The sample variance is a random variable following a scaled χ^2-distribution with $n-1$ degrees of freedom. More specifically,

$$V = (n-1)\frac{s^2}{\sigma^2} \sim \chi^2(n-1) \tag{4.6}$$

The random variables defined in (4.2)/(4.3) and a rearrangement of (4.6) is now substituted into the numerator and denominator of (4.5), respectively, which gives

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \tag{4.7}$$

Recall that under H_0 the standardized sample mean $Z \sim N(0,1)$ (cf. (4.2)) and accordingly T follows a central t-distribution with $n-1$ degrees of freedom. Under H_1, $Z \sim N\left(\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}}, 1\right)$ (cf. (4.3)) and according to (4.5) then T follows a non-central t-distribution with non-centrality parameter

$$\lambda = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \tag{4.8}$$

The distributions of T under H_0 and H_1 are seen in Figure 4.2.

Note that for calculating (4.7) we do not need to know the true variance since it enters both into the numerator and denominator and cancels out. With the price of heavier tails, the t-distribution incorporates the uncertainty about the true variance when the sample variance is used as an estimate. This makes the t-distribution very useful in small sample cases with unknown variance. As $n \to \infty$ the t-distribution approaches the standard normal distribution and the exposition in section 4.2.2 is thus applicable.

From Figure 4.2 it is clear that the power for a one-sided test with significance level α can be obtained by

$$\text{power} = 1 - \beta = 1 - T_{n-1}\left(t_{\alpha,n-1}\middle|\frac{|\mu_1 - \mu_0|}{\sigma/\sqrt{n}}\right)$$

where $T_{n-1}(\cdot|\lambda)$ is the cumulative distribution function of the non-central t-distribution with $n-1$ degrees of freedom, $t_{\alpha,n-1}$ is the critical value obtained from the distribution of T under H_0 and λ is the non-centrality parameter given in (4.8).

Power for at two-sided test with significance level α is obtained by

$$\text{power} = 1 - T_{n-1}\left(t_{\alpha/2,n-1}\middle|\frac{(\mu_1 - \mu_0)}{\sigma/\sqrt{n}}\right) + T_{n-1}\left(-t_{\alpha/2,n-1}\middle|\frac{(\mu_1 - \mu_0)}{\sigma/\sqrt{n}}\right) \quad (4.9)$$

and by ignoring a small value $< \alpha/2$ approximate power for a two-sided hypothesis test is

$$\text{power} = 1 - T_{n-1}\left(t_{\alpha/2,n-1}\middle|\frac{|\mu_1 - \mu_0|}{\sigma/\sqrt{n}}\right) \quad (4.10)$$

[...subsections left out...]

4.2.5 Further continuation of tutorial paper

A natural continuation of the tutorial paper includes power calculations for the most widely used hypothesis tests, comprising two-sample t-tests, paired t-tests and a few variants of ANOVA. The hypothesis tests are in the current draft accompanied by R code, where statistical power are computed both manually in line with the closed-form expressions and by use of relevant R packages. It is also a possibility to cover the use of simulation providing examples illustrating the versatility and simplicity of this approach.

4.3 Statistical power for hierarchical models

The following section deals with statistical power for hierarchical data structured as *in vivo* Comet assay data. First, closed-form expressions for statistical power are derived. Next, the second part of Paper D providing power curves for *in vivo* Comet assay studies on testicular cells is summarized.

The closed-form expressions presented below are valid for balanced data. It is not a limitation in this setting as Comet assay data commonly are balanced, and which not is unusual for designed studies in general. Furthermore, data are assumed to be normally distributed, which based on the examined data seems reasonable when the within-sample distributions are summarized by the median of the log-transformed data (see Section 2.3 on page 13).

4.3.1 Closed-form expressions

The structure of the summarized Comet assay data are outlined in Figure 2.3. A suitable model for this data is a linear mixed-effects model as defined in (2.2) on page 11. The hypothesis of interest is concerning equality of the different treatment groups

$$\begin{aligned} H_0&: \ \tau_1 = \tau_2 = \cdots = \tau_a = 0 \\ H_A&: \ \text{at least one } \tau_i \neq 0 \end{aligned}$$

and closed-form expressions for the statistical power are derived for this specific test.

When a hypothesis test is performed within this framework, we calculate the test statistic as

$$F = \frac{bn \sum_{i=1}^{a} (\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2/(a-1)}{n \sum_{i=1}^{a} \sum_{j=1}^{b} (\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot})^2/(a(b-1))}, \tag{4.11}$$

(Montgomery, 2005) and compare it to a central F-distribution with $a-1$, $a(b-1)$ degrees of freedom to obtain a p-value, i.e. the probability of observing this value or something more extreme given the null hypothesis. To calculate power we also need information on the alternative distribution, which will be assessed in the following.

First, let $y_{ij\cdot} = \sum_{k=1}^{n} y_{ijk}$, $y_{i\cdot\cdot} = \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}$, $y_{\cdots} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}$ and let

$\bar{y}_{ij\cdot} = \frac{1}{n} y_{ij\cdot}$, $\bar{y}_{i\cdot\cdot} = \frac{1}{bn} y_{i\cdot\cdot}$, $\bar{y}_{\cdot\cdot\cdot} = \frac{1}{abn} y_{\cdot\cdot\cdot}$. The observations y_{ijk} and the group averages $\bar{y}_{ij\cdot}$, $\bar{y}_{i\cdot\cdot}$ and $\bar{y}_{\cdot\cdot\cdot}$ are realizations of the random variables Y_{ijk}, $\overline{Y}_{ij\cdot}$, $\overline{Y}_{i\cdot\cdot}$ and $\overline{Y}_{\cdot\cdot\cdot}$, respectively. They are distributed as

$$\begin{aligned} Y_{ijk} &\sim N\left(\mu + \tau_i, \sigma_\beta^2 + \sigma^2\right) \\ \overline{Y}_{ij\cdot} &\sim N\left(\mu + \tau_i, \frac{n\sigma_\beta^2 + \sigma^2}{n}\right) \\ \overline{Y}_{i\cdot\cdot} &\sim N\left(\mu + \tau_i, \frac{n\sigma_\beta^2 + \sigma^2}{bn}\right) \\ \overline{Y}_{\cdot\cdot\cdot} &\sim N\left(\mu, \frac{n\sigma_\beta^2 + \sigma^2}{abn}\right) \end{aligned} \tag{4.12}$$

Further details are given in Appendix A in Paper B. Furthermore,

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\cdot})^2 \sim \sigma^2\chi^2\left(ab(n-1)\right) \tag{4.13}$$

$$n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot})^2 \sim (\sigma_\beta^2 + \sigma^2)\chi^2\left(a(b-1)\right) \tag{4.14}$$

$$bn\sum_{i=1}^{a}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot})^2 \sim (\sigma_\beta^2 + \sigma^2)\chi^2\left(a-1, \lambda\right) \tag{4.15}$$

where

$$\lambda = \frac{bn\sum_{i=1}^{a}\tau_i}{n\sigma_\beta^2 + \sigma^2}$$

See Appendix B in Paper B for details.

Consider the sum of squares related to the treatment effect. As $\overline{Y}_{i\cdot\cdot}$ is a normal random variable then

$$\frac{\sum_{i=1}^{a}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot})^2}{\mathrm{Var}(\overline{Y}_{i\cdot\cdot})} \sim \chi^2(a-1, \lambda),$$

where

$$\lambda = \frac{\sum_{i=1}^{a}(E(\overline{Y}_{i\cdot\cdot}) - E(\overline{Y}_{\cdot\cdot\cdot}))^2}{\mathrm{Var}(\overline{Y}_{i\cdot\cdot})}$$

(Johnson et al., 1995). Inserting from (4.12) and (4.15) yields

$$\frac{bn\sum_{i=1}^{a}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot})}{n\sigma_\beta^2 + \sigma^2} \sim \chi^2(a-1, \lambda), \tag{4.16}$$

where

$$\lambda = \frac{bn \sum_{i=1}^{a} \tau_i^2}{n\sigma_\beta^2 + \sigma^2}.$$

Now, consider the sum of squared due to animals. As above, since $\overline{Y}_{ij\cdot}$ is normally distributed then

$$\frac{\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot})^2}{\mathrm{Var}(\overline{Y}_{ij\cdot})} \sim \chi^2(a(b-1), \lambda),$$

where

$$\lambda = \frac{\sum_{i=1}^{a}(E(\overline{Y}_{ij\cdot}) - E(\overline{Y}_{i\cdot\cdot}))^2}{\mathrm{Var}(\overline{Y}_{ij\cdot})}.$$

According to (4.12) and (4.14) then

$$\frac{n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot})^2}{n\sigma_\beta^2 + \sigma^2} \sim \chi^2(a(b-1)), \tag{4.17}$$

since

$$\lambda = \frac{\sum_{i=1}^{a}((\mu+\tau_i) - (\mu+\tau_i))^2}{\mathrm{Var}(\overline{Y}_{ij\cdot})} = 0.$$

The ratio of two independent χ^2-distributed random variables, $V_1 \sim \chi^2(\nu_1, \lambda_1)$ and $V_2 \sim \chi^2(\nu_2, \lambda_2)$, each divided by their corresponding degrees of freedom follows an F-distribution, in that

$$\frac{V_1/\nu_1}{V_2/\nu_2} \sim F(\nu_1, \nu_2, \lambda_1, \lambda_2)$$

(Johnson et al., 1995). It can be shown using Fisher-Cochran's Theorem (Rao, 1973) that (4.16) and (4.17) are independent, and we then have information on the distribution of their ratio when they each is divided by their degrees of freedom

$$F = \frac{bn\sum_{i=1}^{a}\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right)^2/(a-1)}{n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot}\right)^2/(a(b-1))} \sim F\left(a-1, a(b-1), \lambda\right), \tag{4.18}$$

where

$$\lambda = \frac{bn \sum_{i=1}^{a} \tau_i^2}{n\sigma_\beta^2 + \sigma^2}. \tag{4.19}$$

The ratio in (4.18) is in fact the sample statistic that is defined in (4.11). Along with information on its distribution it forms the basis for the inference to be made.

The cumulative distribution function of F is denoted $G(\,\cdot\,; \nu_1, \nu_2, \lambda)$, and the critical value, $F_{\alpha;\nu_1,\nu_2}$ is the $(1-\alpha)$th quantile of the null distribution, such that $G(F_{\alpha;\nu_1,\nu_2}; \nu_1, \nu_2) = 1 - \alpha$. Under the null hypotheses we have that $\tau_i = 0,\ i = 1, ..., a$ so that $\lambda = 0$, i.e. the F-statistic in (4.18) follows a central F-distribution.

The statistical power is the probability of correctly rejecting the null hypothesis when it indeed is false, and can thus be formulated as

$$\text{Power} = 1 - G(F_{\alpha;a-1,a(b-1)}; a-1; a(b-1), \lambda), \tag{4.20}$$

where λ is given in (4.19).

When assessing statistical power prior to carrying out the actual study, it is often difficult to specify values for all τ_i. One way to overcome this is to specify the difference in treatment means of the groups with the lowest and highest response, D (Montgomery, 2005). The price we pay is that all remaining groups are assumed to have an average response, implying that $\tau_{i'} = 0, i' = 1, ..., a-2$, and these groups do not add anything to the non-centrality parameter. The power calculations will thus give the *minimum* statistical power for the specified value of D, in which case the non-centrality parameter is given as

$$\lambda = \frac{bn \sum_{i=1}^{2} (\frac{1}{2}D)^2}{n\sigma_\beta^2 + \sigma^2} = \frac{bnD^2}{2(n\sigma_\beta^2 + \sigma^2)} \tag{4.21}$$

4.3.2 Power and sample size for Comet assay studies

The results given in this section summarizes the second part of Paper D. It builds on the results outlined in Section 2.3 on page 13, where it was concluded that an appropriate summary statistic for *in vivo* Comet assay data on testicular cells is the median of the log-transformed data.

The summarized data that were collected in this study were fitted model (2.2) defined on page 11. According to the Comet assay literature it is common prac-

tice to score 50 or 100 cells per slide and we provided power curves accordingly. When 50 cells were scored per slide the estimated variance components were

$$\hat{\sigma}^2_\beta = 0.09, \;\; \hat{\sigma}^2 = 0.11.$$

When 100 cells were scored per slide then

$$\hat{\sigma}^2_\beta = 0.08, \;\; \hat{\sigma}^2 = 0.09.$$

These variance components form the basis of the power curves that are seen in Figure 4.3. The first column outlines the power curves when 50 cells are sampled per slide, while the second column is valid when 100 cells are sampled per slide. Due to the log-transformation the effect sizes are given as fold changes rather than differences. The fold change corresponds to the difference, D, as described above, so that the fold change is specified for the maximum response mean relative to the minimum response mean.

It is not uncommon to use between 3 and 8 animals in each group in Comet assay studies. It is seen from the power curves that a fairly large fold change is actually required to obtain power at the conventional level at 80% with a sample size in that range.

The power curves are intended to be used by biomedical researchers in the design of prospective Comet assay studies. Further reflections on the power curves that are relevant for this audience are given in Paper D.

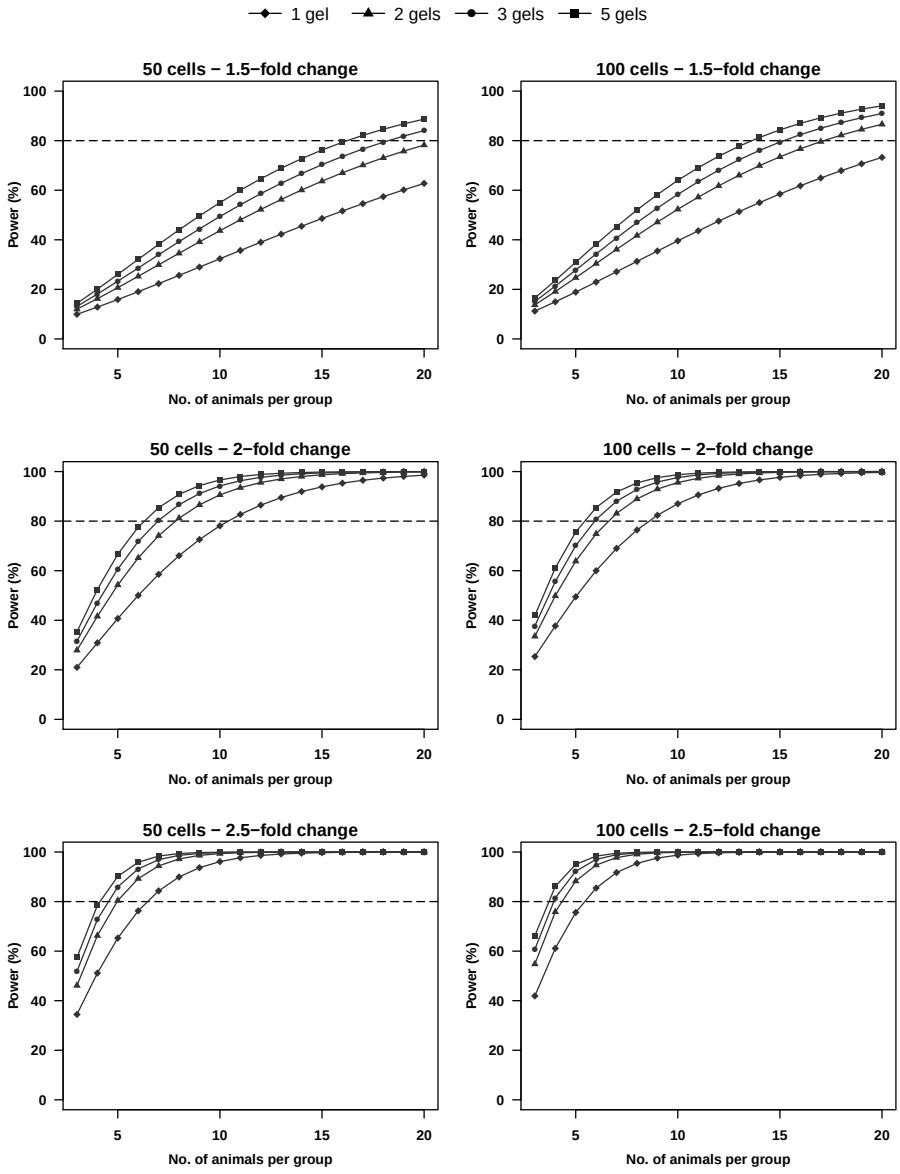

Figure 4.3: Power curves outlining the number of animals per group and gels per animal required to detect certain fold changes with a power of 80% (dotted line) when 50 cells (first column) and 100 cells (second column) are scored per gel, respectively. The power calculations apply when the summary statistic is the median of the log-transformed data.

Chapter 5

Agreement studies

The objective of some biomedical studies is to evaluate the degree of agreement between different assessors. These assessors can range from competing medical instruments and assays, over doctors and technicians, to everyday consumers. Such studies are called agreement studies and can have several aims. They can include the validation of a new diagnostic assay against a gold standard or they may concern how well doctors agree in diagnosing patients. Whereas agreement studies can involve humans both on the assessor and the subject level they can also include non-humans such as assays, samples and animals at both of these levels.

Agreement studies are quite crucial in the validation of the quality of data emerging from biomedical studies. While Chapter 3 and 4 of this thesis have been dealing with issues related to the statistical models, the current chapter is concerned with the appraisal of the quality of collected data. First, general aspects regarding agreement studies are presented. Next, Paper E dealing with evaluation of a national drug related problems database is summarized.

5.1 Background

Barnhart et al. (2007) describes *agreement* as a measure of "closeness" between

readings. Moreover, the US Food and Drug Administration (FDA) defines *accuracy* as "the degree of closeness of the determined value to the nominal or known true value under prescribed conditions" and *precision* as "the closeness of agreement (degree of scatter) between a series of measurements obtained from multiple sampling of the same homogenous sample under the prescribed conditions" (Food and Drug Administration (FDA), 2001). Barnhart et al. (2007) and Lin (2007) assert that agreement encompasses both accuracy and precision. Accuracy and precision are concepts that are related to the terms bias and variance, respectively, that in some fields are more commonly used. In case of disagreement it is useful to examine whether it is caused by inaccuracy (bias) or stems from random variation. The former is often attributed to a calibration problem and is easier to fix than the latter (Lin, 2007).

In agreement studies there are multiple recordings on the same subject, that is, a number of subjects are each rated by a number of assessors. This can be handled by a number of statistical models, yet agreement studies are treated somewhat separately in the literature. This may be attributable to that the main interest often lies in making inference on a measure quantifying the *degree of agreement* between the different assessors (Nawarathna and Choudhary, 2013). This measure can be estimated directly or it can be a function of parameters of a fitted model.

A number of different factors impact how data from agreement studies are properly summarized, analyzed and interpreted. Most importantly, the response variable can be classified according to its measurement scale, which includes a nominal, ordinal or continuous scale. Also, a "gold standard" (the best test available (Versi, 1992)) can be present or absent and validity studies often aim to compare new methods or raters to a gold standard. Furthermore, the number of assessors is of importance with the most significant difference between two and multiple assessors. In case of a nominal response variable the number of categories are of importance with the main difference being between two and multiple categories. The next section is concerned with cases where the response is on a nominal scale.

5.1.1 Nominal endpoint

In many fields the kappa statistic is a very popular measure for summarizing agreement data on a nominal scale. The first kappa-like statistic was introduced in 1892 by Galton (Galton, 1892; Smeeton, 1985). Another related statistic is Scott's pi, which was presented in 1955 (Scott, 1955). Cohen proposed the kappa statistic in 1960 (Cohen, 1960) that has become very popular and still is widely used in many research areas. A number of modified related statistics were later

suggested (see e.g. Cohen, 1968; Conger, 1980; Krippendorff, 1980; Brennan and Prediger, 1981; Schuster, 2005; Gwet, 2008); one of these is Fleiss' kappa (Fleiss, 1971). Cohen's kappa measures the agreement between two assessors and Fleiss intended to generalize Cohen's kappa to accommodate three or more assessors. However, for two assessors Fleiss' kappa does not reduce to Cohen's kappa but to Scott's pi (Conger, 1980).

In some fields Fleiss' kappa is the summary measure of choice when there are more than two assessors. It seems to be a common misperception that Fleiss' kappa is a generalization of Cohen's kappa (which in part may be due to that Fleiss named his statistic kappa and not pi) and this may explain the common use. Contrary to Cohen's kappa, Fleiss' kappa allows different assessors to rate different subjects (Fleiss, 1971).

Fleiss' kappa is given as

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \tag{5.1}$$

where $\overline{P}$ is the overall proportion of observed agreement and $\overline{P}_e$ is the proportion of agreement that would be expected if the ratings were assigned purely by chance (Fleiss, 1971). Fleiss' kappa is thus the ratio of observed agreement that exceeds agreement by chance to the agreement that is attainable beyond chance. More specifically, let N be the total number of subjects, let n be the number of ratings per subject, let k be the number of categories that can be assigned to the subjects and let n_{ij} be the number of raters who categorized the ith subject to the jth category, where $i = 1, ..., N$ and $j = 1, ..., k$. The proportion of agreeing pairs of raters out of all possible pairs of raters is

$$\begin{aligned} P_i &= \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \\ &= \frac{1}{n(n-1)} \left(\sum_{j=1}^{k} n_{ij}^2 - n \right) \end{aligned} \tag{5.2}$$

The overall proportion of observed agreement is the average of the P_i's

$$\begin{aligned} \overline{P} &= \frac{1}{N} \sum_{i=1}^{N} P_i \\ &= \frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - Nn \right) \end{aligned} \tag{5.3}$$

The proportion of all assignments to the jth category is

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \tag{5.4}$$

and the expected agreement purely by chance is given as

$$\overline{P}_e = \sum_{j=1}^{k} p_j^2 \tag{5.5}$$

The reasoning behind the calculation of the expected agreement, $\overline{P}_e$, was not made explicit by Fleiss.

Despite the popularity of the different kappa indices they are surrounded by quite some controversy and they have repeatedly been objects of debate (see e.g. Feinstein and Cicchetti, 1990; Sim and Wright, 2005; de Mast, 2007). A thorough introduction to these issues is also provided by John Uebersax[1]. One concern is that the same observed proportion of agreement can be associated with fairly different kappa values depending on the marginal distributions and this seems to be the main point of criticism associated with the kappa statistic. Suggested premises concerning the expected agreement, $\overline{P}_e$, are given by de Mast (2007). If these premises hold, then for each rater the probability distribution of chance ratings is assumed to be equal to the observed marginal distribution. Accordingly, with an uneven distribution of the marginal totals, agreement by chance will happen more frequent (due to the categories with a high prevalence), hence $\overline{P}_e$ increases in turn implying that κ decreases. It has been suggested to assume a uniform distribution of the chance ratings, which resolves some of the issues associated with Fleiss' kappa (Brennan and Prediger, 1981; de Mast, 2007).

A number of alternative statistical approaches which are based on explicit model formulations have been suggested for agreement data. One approach is latent structure analysis or latent variable analysis which covers different methods including latent class analysis and latent trait analysis. Latent structure models assume that the observed rating depend on latent (unobserved) variables. Latent class models are useful for discrete latent variables, whereas latent trait models assume a continuous latent variable (Uebersax, 1992). Other related approaches are log-linear models, association models and quasi-symmetry models (Uebersax, 1992; Agresti, 2013). One advantage is that possible structures and covariates can be identified and appropriately included in the models. However, it seems as these model-based approaches have not gained footing in many scientific areas and the kappa-like indices still prevail. This may be due to a common practice that has evolved within some research fields. Also, the kappa indices are fairly easy to grasp, calculate and interpret (or at least they seem to be; c.f. the concerns mentioned earlier) compared to the model-based approaches.

[1] http://www.john-uebersax.com/stat/agree.htm

5.2 Evaluation of a drug related problems database

The topic of paper E is the evaluation of a database containing registrations of drug related problems (DRPs) at hospitals in Denmark. Suboptimal medication treatment can lead to increased morbidity, mortality and costs and one approach to improving the medication treatment is to let pharmacists conduct medication reviews to identify drug related problems (DRPs). Changes to the medication treatment are then recommended based on the reported DRPs. To document and standardize DRPs identified by clinical pharmacists at the hospitals in Denmark a DRP database was developed and it was implemented July 1, 2010. Three years after the implementation around 125,000 registrations were made and the comprehensive DRP database provides a possibility to conduct national analyses. With this in view the aim of this study was to evaluate use in practice, reliability and reproducibility of the Danish DRP database.

The DRPs are assigned to one of the following 14 categories: dose, dosing time and interval, side effect, interaction, drug form and strength, compatibility, non-adherence to guidelines, therapeutic duplication, drug allergies, length of treatment, supplement to treatment, electronic patient chart related, inappropriate drug or no DRP. In practice, the categories compatibility and no DRP are rarely used and are not considered further.

A project group consisting of 7 clinical pharmacists designed and supervised the study. It contains three substudies, of which I was involved in the last two; hence these two substudies will be outlined in the following sections.

The aims of this study are to assess the reliability (substudy two) and the reproducibility (substudy three) of the DRP database. In both cases the objectives are quantified by assessing the agreement between different assessors. As mentioned above there are different approaches to evaluate such agreement. While the various versions of the kappa statistics are associated with some concerns they have become the *de facto* standard in agreement studies within various different fields. The large number of categories (12 categories) in the current study combined with the observed marginal distributions gives rise to an expected agreement of an inconsiderable magnitude implying that Fleiss' kappa and the observed proportion agreement in all cases in the current study are similar. The main issues of Fleiss' kappa are thus resolved in this particular case. Because of this we find it sensible to follow common practice within the target research field and to quantify the results by means of Fleiss' kappa. It is accompanied by other relevant measures such as observed proportions of agreement and marginal distributions as these also convey valuable information.

Table 5.1: Examples of DRPs and recommendations categorized by the project group and the respondents

Description of DRP	Recommendation	Categorization according to the project group	Number of respondents with similar categorization
The patient is treated with IV cefuroxime and IV metronidazole. The patient has been treated with IV metronidazole for 1 day and night	The bioavailability for oral metronidazole is almost 100%, hence it is recommended to consider substitute IV to oral treatment	Drug form and strength	34 (100%)
P-potassium is 3.1 mmol/L (reference interval: 3.5-4.6 mmol/L)	Recommendation of potassium supplement	Supplement to treatment	34 (100%)
A patient of 83 years is treated with clopidogrel, aspirin and ibuprophene	Recommendation of ceasing NSAID due to old age and other factors, which increase the risk of GI bleeding	Inappropriate drug	21 (62%)
The patient has an active prescription initiated four years ago of Tobradex eyedrops, susp. 1+3 mg/ml, 1 drop x 3 daily in the right eye	Please consider whether the prescription of Tobradex should be active. If the patient has no current need for the prescription, please consider ceasing it	Length of treatment	10 (29%)

5.2.1 Substudy two: inter-rater reliability study

The second substudy of Paper E concerns *inter-rater reliability*. The project group identified 24 cases of DRPs and recommendations that were recategorized by 34 regular users of the DRP database. In addition, gold standards (GS) for the ratings of these cases were appointed by the project group. Some examples are listed in Table 5.1.

The inter-rater agreement was quantified with observed proportion of agreement, $\hat{p}_o$ and Fleiss' kappa, $\hat{\kappa}$. Also, the proportion of agreement with the gold standard, $\hat{p}_{GS}$, was calculated. The observed agreement and Fleiss' kappa are measures of the agreement among the raters (not including the GS) whereas $\hat{p}_{GS}$ is a measure of the agreement of the raters with GS. Bootstrap confidence intervals (CI) for all three measures were obtained using the bias-corrected accelerated (BC_A) method (Efron, 1987; Hall, 1988) based on 10,000 bootstrap replicates.

The overall agreement of the 34 clinical pharmacists was $\hat{p}_o = 0.81$ with 95% CI (0.72; 0.89) and Fleiss kappa was $\hat{\kappa} = 0.79$ with 95% CI (0.70; 0.88). Comparing each rater to GS gave kappa values all greater than 0.6 and half of them exceeded 0.8. The agreement between the raters and the GS are depicted in Figure 5.1. For 7 of the 12 categories the proportion of agreement of the 34 clinical pharmacists with GS exceeded 0.90 whereas for 2 additional categories the agreement was above 0.80. The overall agreement of each rating with GS was $\hat{p}_{GS} = 0.81$ with 95% CI (0.68; 0.88). The categories with the lowest degree of agreement with GS were "side effects", "length of treatment" and "inappropriate drug".

5.2.2 Substudy three: reproducibility study

Two members of the project group re-categorized a random sample of existing records from the DRP database. The project group members re-categorized 379 records based on the text field, and they were blinded to the initial categorization.

The observed agreement and Fleiss' kappa was calculated both overall and specific to each category. In addition, the observed agreement and Fleiss' kappa was calculated for the DRP database and the two project group members. Bootstrap confidence intervals (CI) for all measures were obtained with the bias-corrected accelerated (BC_A) method based on 10,000 bootstrap replicates.

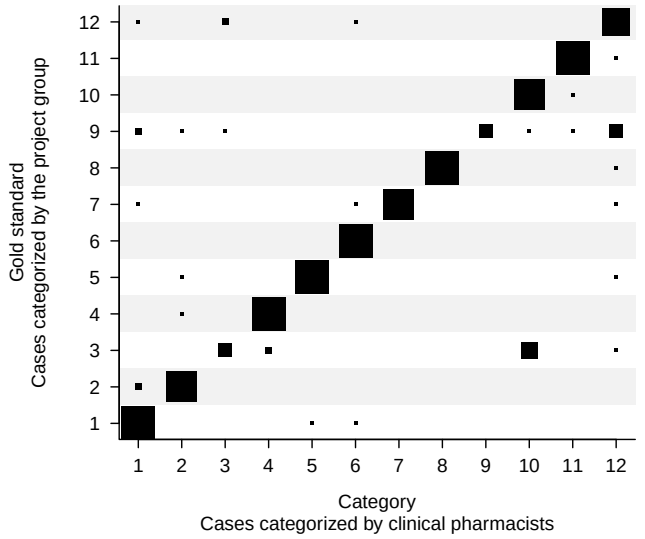

Figure 5.1: Categorization of 24 cases by the individual raters and the project group (gold standard). The x-axis specifies the categories made by the clinical pharmacists and the y-axis is the gold standard, i.e. the categorization made by the project group. Each square represents the proportion agreement of the clinical pharmacists with the gold standard (calculated for each category of the gold standard), implying that the sizes of the squares sum to 1 horizontally.

The overall observed agreement was $\hat{p}_o = 0.83$ with 95% CI (0.80; 0.86) and Fleiss' kappa was $\hat{\kappa} = 0.81$ with 95% CI (0.78; 0.85). The pairwise agreement between the categories documented in the DRP database and the categorizations by the project group members (PGM) were

- DRP database and PGM1
  $\hat{p}_o = 0.83$ with 95% CI (0.78; 0.86) and $\hat{\kappa} = 0.81$ with 95% CI (0.76; 0.85)

- DRP database and PGM2
  $\hat{p}_o = 0.85$ with 95% CI (0.81; 0.88) and $\hat{\kappa} = 0.81$ with 95% CI (0.79; 0.87)

- PGM1 and PGM2
  $\hat{p}_o = 0.82$ with 95% CI (0.78; 0.86) and $\hat{\kappa} = 0.80$ with 95% CI (0.76; 0.84)

Category-specific agreement measures and marginal distributions are given in Table 5.2. The categories are associated with varying degree of agreement although the majority of the categories are consistently categorized by the DRP and the two project group members, PGM1 and PGM2. The kappa values for 7 out of 12 categories are above 0.80 and all kappa values are greater than 0.60. For the cases where the assessors disagreed there were no consistent alternative categorizations for any of the categories.

5.2.3 Concluding remarks

Two aims of the current study were to evaluate the inter-rater reliability and the reproducibility of a national DRP database. Due to the number of categories and the observed marginal distributions we considered it sensible to use Fleiss' kappa in line with common practice of the target research field.

All results show a high degree of agreement between the different assessors. This indicates that the entries in the DRP database are of sufficient quality to conduct ensuing national analyses.

Table 5.2: Marginal distributions (distribution of rated categories for each of DRP and the two project group members, PGM1 and PGM2) and category-specific agreement measures with 95% confidence intervals (CI). For the category "Drug allergies" all raters agreed and no bootstrap confidence intervals could be obtained.

	Marginal distributions			Observed agreement		Fleiss' kappa	
Rated category	DRP	PGM1	PGM2	Estimate	CI	Estimate	CI
Dose	0.16	0.16	0.18	0.86	(0.80; 0.91)	0.83	(0.76; 0.89)
Dosing time and interval	0.11	0.08	0.09	0.84	(0.75; 0.90)	0.82	(0.73; 0.89)
Side effect	0.03	0.04	0.02	0.61	(0.40; 0.79)	0.60	(0.38; 0.79)
Interaction	0.05	0.06	0.05	0.80	(0.65; 0.89)	0.79	(0.63; 0.88)
Drug form and strength	0.03	0.02	0.03	0.87	(0.69; 0.97)	0.87	(0.68; 0.97)
Non-adherence to guidelines	0.15	0.17	0.16	0.92	(0.87; 0.95)	0.90	(0.84; 0.94)
Therapeutic duplication	0.05	0.05	0.06	0.89	(0.77; 0.96)	0.88	(0.76; 0.95)
Drug allergies	0.03	0.03	0.03	1.00	-	1.00	-
Length of treatment	0.03	0.05	0.03	0.70	(0.52; 0.85)	0.69	(0.50; 0.84)
Supplement to treatment	0.14	0.13	0.14	0.88	(0.82; 0.93)	0.87	(0.79; 0.92)
Electronic patient chart related	0.08	0.09	0.08	0.77	(0.66; 0.85)	0.75	(0.63; 0.84)
Inappropriate drug	0.13	0.12	0.13	0.72	(0.63; 0.80)	0.68	(0.58; 0.76)

Chapter 6

Diagnostics on binomial regression models

Observations from some biomedical studies each take one of two possible forms, e.g. the response are categorized as dead or alive, yes or no, agree or disagree or, by convention, as success or failure. Such data are said to be binary. In some cases the observations can be grouped according to some factors, i.e. when a number of subjects have been exposed to the same dose of a compound. In such cases the response can be summarized as the proportion of successes out of the total number of subjects in that particular group. Grouped binary data are called binomial data.

Logistic regression is a popular approach for analyzing binary and binomial data and is widespread in many research areas, in particular biomedicine. Logistic regression models belong to a class of models known as generalized linear models, which accommodate response variables having distributions that are members of the exponential family; two of these are the normal and the binomial distribution. The idea of the generalized linear model is that the response variable is linearly related to the explanatory variables via a specified link function. For logistic regression the link function is the logit link; other common link functions are the probit and the complementary log-log link function. We will refer to *binomial regression models* as generalized linear models with a binary or binomial response variable and appropriate link function.

All statistical models make assumptions and prior to making inference it is important to assess if the model assumptions are satisfied. Numerous approaches for classical linear models assuming normality have been developed and are implemented in common statistical software packages. For non-normal models we however find the range of diagnostic tools to be somewhat more limited. Some functions relevant for performing diagnostics on binomial regression models are available in the statistical software package R; some of these are provided in base R while others are implemented in packages such as car (Fox and Weisberg, 2011) and rms (Harrell Jr, 2014). However, a number of tools that we find useful for evaluating model validity were yet to be implemented. This served as a motivation to develop the R package binomTools that provides diagnostic tools for binomial regression models. The package is publicly available at the CRAN repository.

6.1 The binomTools package

The R package binomTools provides the following functionalities:

Data sets: The data sets `beetles` and `serum` are supplied to illustrate the use of the functionality available in binomTools.

Residuals: Exact deletion residuals are implemented in `exact.deletion`. A wrapper function `Residuals` seeks to add clarity on which residual type that is returned.

Graphical diagnostic tools: A half normal plot is implemented in `halfnorm`. Profile likelihoods for the parameters in the binomial regression model are obtained with `profile`. The R^2 measure *coefficient of discrimination* as proposed by Tjur (2009) is implemented in `Rsq`†.

Goodness-of-fit tests: The `HLtest`† and `X2GOFtest`† are implementations of the Hosmer-Lemeshow and Pearson's χ^2 goodness-of-fit tests, respectively.

Convenience functions: The empirical logit transform can be calculated with `empLogit`†. The observations can be grouped according to the covariate structure with `group`†.

The package documentation is given in Appendix F. Functions marked with † was implemented by a co-author and are not considered further in this thesis. The remaining functions are outlined in the following.

6.1.1 Residuals

Upon model fitting the residuals can be calculated in various ways. Different residual types can each add insight into the validity of the model and several definitions therefore exist. Unfortunately, the residual terminology seems to be somewhat confused. Some residuals go by different names and the same term can refer to different residual types. Even for the classical linear model this is an issue and one example concerns the *Studentized* residuals. This term is commonly used, and some authors distinguish between internal Studentization and external Studentization (Margolin, 1977; Cook and Weisberg, 1982; Gray and Woodall, 1994). However, Cook (1977) and Cook and Weisberg (1982) refer to the internal Studentized residuals as Studentized residuals while they by many others are called standardized residuals (Hoaglin and Welsch, 1978; Belsley et al., 1980; Atkinson, 1985; Fox and Weisberg, 2011). External Studentized residuals are also sometimes called Studentized residuals (Hoaglin and Welsch, 1978; Belsley et al., 1980) as well as cross-validatory or jack-knife (Atkinson, 1981), deletion (Atkinson, 1985) and Studentized deleted (Dupuis and Hamilton, 2000) residuals.

In the non-normal case the residuals may be extended in different ways and this adds to the confusion. Examples of residuals defined for generalized linear models are raw residuals, response residuals, working residuals, partial residuals, Pearson residuals, standardized Pearson residuals, deviance residuals, standardized deviance residuals, adjusted deviance residuals, likelihood residuals, score residuals, Anscombe residuals, standardized Studentized residuals, modified standardized Pearson residuals, Cox-Snell residuals etc. (Fox, 2002; Collett, 2003; Hardin and Hilbe, 2007). The number of residual types combined with the confused terminology may hinder the perspective. Many authors seem to be aware of the problem and carefully define the residuals that they refer to. However, we do not find the documentation on the different residual functions in R very explicit, and we believe that it considerably impedes the exact understanding of which residuals that are assessed with the functions available in R. This may in particular hold true for novel users.

In binomTools we implemented the function `Residuals` that seeks to add clarity on the residual types that are returned from relevant functions in R. The `Residuals` function is thus a wrapper function that calls appropriate functions given an argument specified by the user. We sought to provide a thorough documentation that adds clarity on this matter. All residual types are defined in the documentation given in Appendix F.

In addition to the residuals available in base R we implemented `exact.deletion` that provides exact deletion residuals. The ith deletion residual is calculated

by subtracting the deviances when a binomial regression model is fitted the full set of n observations and the same model is fitted to a set of $n-1$ observations excluding the ith observation, $i = 1, ..., n$. These residuals are also known as likelihood residuals, Studentized residuals, externally Studentized residuals, deleted Studentized residuals and jack-knife residuals.

6.1.2 Half normal plot

Suitable residuals of a binomial regression model, such as the standardized deviance residuals or likelihood residuals, can be reasonably approximated by a normal distribution given that the fitted model is valid and the binomial denominators are reasonably sized. Still, the residuals may deviate from an exact straight line in a normal plot, especially when the binomial denominators are small. A half normal plot with simulated envelopes (Atkinson, 1981) may therefore prove useful, in particular for detecting model inadequacy and outliers. We implemented the half normal plot for binomial regression models as exposited in Collett (2003). A half-normal plot uses the absolute residual values but is otherwise equivalent to a normal plot. The simulated envelopes are constructed such that, for each of the n observations, additional observations are simulated from the model and the minimum, average and maximum values are outlined on the half-normal plot. The envelopes thus indicate the boundaries where the residuals are likely to fall given that the model is correct. The half normal plot is optionally interactive so that a number of residuals can be identified by clicking the points in the plot. An example of the half normal plot with simulated envelopes is seen in Figure 6.1.

6.1.3 Profile likelihoods

Upon model fitting it may be valuable to examine the profile likelihood for the parameters that enter the binomial regression model. The MASS package (Venables and Ripley, 2002) provides profile likelihoods for parameters of a generalized linear model and the appertaining plot function provides plots of the profile likelihood root for each of these parameters. We provided a new plot method in binomTools, which extends the plot functionality of the profile likelihoods. Some of the most important extensions are the option to plot the profile likelihood on a relative or absolute scale instead of the profile likelihood root and to add a quadratic approximation to assess the regularity of the profile likelihoods.

An example of a profile likelihood plot is given in Figure 6.2. In this case

the 95% Wald CI for the second parameter (`dose0.045`) shown to the right is $(-7.8; -3.4)$ while the corresponding profile likelihood CI is $(-8.6; -3.8)$. This is consistent with the profile likelihood and the quadratic approximation that are shown in Figure 6.2.

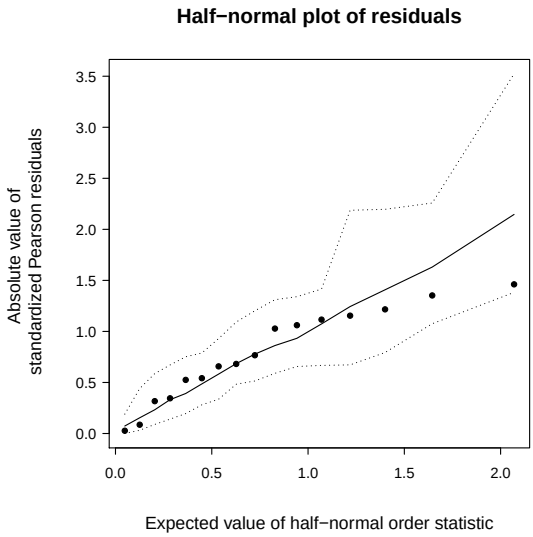

Figure 6.1: An example of a half normal plot with simulated envelopes produced by binomTools.

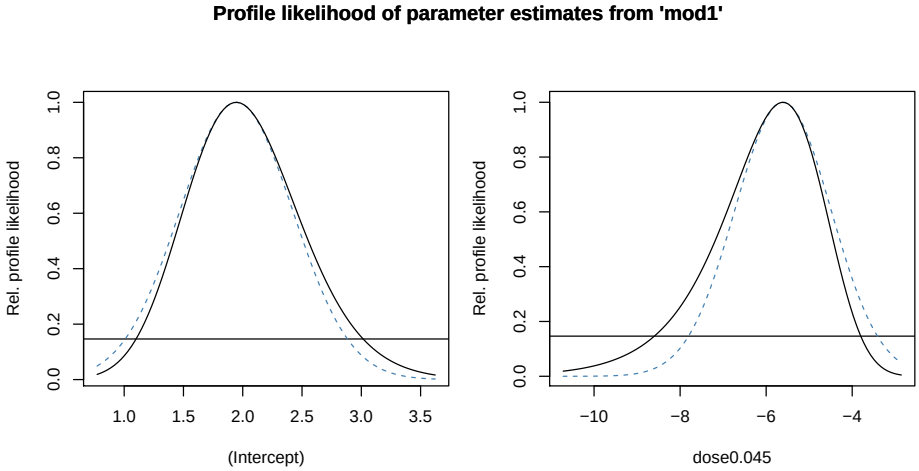

Figure 6.2: The profile likelihood for two parameters of a binomial regression model. From this plot the regularity of the profile likelihoods can be examined.

Chapter 7

Concluding remarks

Innumerable biomedical studies are continuously conducted, through which results that are of pertinence to most people are accomplished. This thesis is dealing with the design and analysis of biomedical studies. In this chapter the main conclusions from the thesis are highlighted and briefly discussed.

Statistical analysis of *in vivo* Comet assay data. In Paper D we identify the median of the log-transformed data as the most suitable statistic to summarize *in vivo* Comet assay testicular data out of 23 candidate summary statistics. We propose that the summarized data are analyzed with a linear mixed-effects model. Our conclusions are based on how well the summarized data comply with the model assumptions.

As Comet assay data are positively skewed, statistical methods assuming normality cannot be applied directly. The question of how to analyze Comet assay data has previously been raised. Different end points are in use though and existing studies on this topic concern other end points than the % tail DNA. Moreover, the proposed summary statistics are in these studies evaluated according to other criteria, e.g. to maximize power (Duez et al., 2003; Wiklund and Agurell, 2003). However, we find it crucial to seek fulfillment of the model assumptions and we consider other criteria, such as increased power, to be of secondary importance.

Type I errors. A literature study presented in Paper B reveals that the hierarchical structure of *in vivo* Comet assay data too often are neglected in the statistical analysis. We formulated closed-form expressions for the approximate type I error rate and assessed the type I error rates for different factor combinations as they appear in practice. The type I error rates are severely inflated in all cases. Papers A and B address these concerns within the framework of *in vivo* Comet assay studies whereas Paper C is directed at a broader audience. Our literature study moreover revealed that the statistical methodology is not thoroughly documented in the examined papers thus leaving doubt about how to properly interpret the reported results.

Our finding in the literature study is supported by other papers (e.g. Strasak et al., 2007; Baccaglini et al., 2010) that report this type of misuse to occur in biomedical studies. We initiated this study to emphasize the gravity in case of this assumptional violation. We find the inflated type I error rate to be problematic for several reasons. First, we consider it likely that researchers are inspired by the statistical methodology and wording in publications within their field and future publications may thus suffer from the same flaws. Second, new studies may be initiated based on a significant finding. When the type I error rate is inflated it may thus lead to waste of resources. For ethical reasons this is especially problematic in case of animal experiments, which includes *in vivo* Comet assay studies. Third, as studies demonstrating a significant effect are more likely to be published, we are concerned that studies that fail to account for the hierarchical structure are over-represented in the literature leading to that evidence for a given compound erroneously accumulates.

Type II errors. Many publications within the biomedical field do not report the use of power calculations prior to conducting the study. Power curves for *in vivo* Comet assay studies on testicular cells are provided in Paper D to guide researchers in the design of future studies. From these power curves the number of animals and the number of slides per animal to use can be obtained.

Sample size recommendations have been provided by Smith et al. (2008) for samples of rat liver, blood, bone marrow and stomach samples. To our knowledge no recommendations for testicular cell samples have been published previously. We find the literature covering power and sample size calculations to be limited, which may impede the understanding of the general concept in turn hindering the recognition of its importance. We hope that our paper on this topic can motivate biomedical researchers to reflect on the sample size when future *in vivo* Comet assay studies are designed.

Diagnostics on binomial regression models. After model fitting it is important to assess if the model assumptions are fulfilled as spurious conclusions otherwise may be drawn. We implemented a range of diagnostic tools for bi-

nomial regression models in the R package binomTools. The package documentation is given in Appendix F. The package is publicly available at the CRAN repository.

Bibliography

Agresti, A. (2013). *Categorical Data Analysis* (third ed.). Wiley Series in Probability and Statistics. New Jersey, NJ: John Wiley & Sons.

Agresti, A. and B. A. Coull (1998). Approximate is better than "Exact" for interval estimation of binomial proportions. *The American Statistician 52*(2), 119–126.

Algina, J. and S. Olejnik (2003). Conducting power analyses for anova and ancova in between-subjects designs. *Evaluation & the Health Professions 26*, 288–314.

Altman, D. G. (1981). Statistics and ethics in medical research. *BMJ 282*(6257), 44–47.

Altman, D. G. and J. M. Bland (1995). Absence of evidence is not evidence of absence. *British Medical Journal 311*(7003), 485–485.

Anderson, K. (2014). *gsDesign: Group Sequential Design.* R package version 2.8-8.

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika 68*(1), 13–20.

Atkinson, A. C. (1985). *Plots, Transformations and Regression.* Oxford Statistical Science Series. New York: Oxford University Press.

Baccaglini, L., J. J. Shuster, J. Cheng, D. W. Theriaque, V. J. Schoenbach, S. L. Tomar, and C. Poole (2010). Design and statistical analysis of oral medicine studies: common pitfalls. *Oral Diseases 16*(3), 233–241.

Baldwin, S. A., D. M. Murray, and W. R. Shadish (2005). Empirically supported treatments or type I errors? Problems with the analysis of data from group-administered treatments. *Journal of Consulting and Clinical Psychology 73*(5), 924–935.

Ball, R. (2012). *ldDesign: Design of experiments for detection of linkage disequilibrium*. R package version 2.0-1.

Barnhart, H. X., M. J. Haber, and L. I. Lin (2007). An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics 17*(4), 529–569.

Belsley, D. A., E. Kuh, and R. E. Welsch (1980). *Regression Diagnostics*. USA: John Wiley & Sons.

Bolker, B. M. (2008). *Ecological Models and Data in R*. Princeton University Press.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics 25*(2), 290–302.

Brennan, R. L. and D. J. Prediger (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement 41*, 687–699.

Bright, J., M. Aylott, S. Bate, H. Geys, P. Jarvis, J. Saul, and R. Vonk (2011). Recommendations on the statistical analysis of the comet assay. *Pharmaceutical Statistics 10*, 485–493.

Brown, M. B. and A. B. Forsythe (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association 60*(346), 364–367.

Buchner, A., E. Erdfelder, F. Faul, and A.-G. Lang (2014). G*Power 3.1.9.2. Retrieved May 29, 2014, from http://www.gpower.hhu.de.

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience 14*(5), 365–376.

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage 63*, 289–300.

Champely, S. (2012). *pwr: Basic functions for power analysis*. R package version 1.1.1.

Chow, S.-C., J. Shao, and H. Wang (2002). A note on sample size calculation for mean comparisons based on noncentral t-statistics. *Journal of Biopharmaceutical Statistics 12*(4), 441–456.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20*(1), 37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin 70*(4), 37–46.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (Second ed.). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin 112*(1), 155–159.

Cohen, P. (2005). Cohen, Jakob. In B. S. Everitt and D. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*. Wiley.

Collett, D. (2003). *Modelling Binary Data* (Second ed.). USA: Chapman & Hall/CRC.

Collins, A. R. (2004). The comet assay for dna damage and repair - principles, applications, and limitations. *Molecular Biotechnology 26*(3), 249–261.

Collins, A. R., V. L. Dobson, M. Dušinská, G. Kennedy, and R. Štětina (1997). The comet assay: what can it really tell us? *Mutation Research 375*, 183–193.

Collins, A. R., M. Dušinská, C. M. Gedik, and R. Štětina (1996). Oxidative damage to dna: Do we have a reliable biomarker? *Environmental Health Perspectives 104*, 465–469.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin 88*(2), 322–328.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics 19*(1), 15–18.

Cook, R. D. and S. Weisberg (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics* (First ed.). London: Chapman and Hall.

de Mast, J. (2007). Agreement and kappa-type indices. *The American Statistician 61*(2), 148–153.

Descôteaux, J. (2007). Statistical power: An historical introduction. *Tutorials in Quantitative Methods for Psychology 3*(2), 28–34.

Duez, P., G. Dehon, A. Kumps, and J. Dubois (2003). Statistics of the comet assay: a key to discriminate between genotoxic effects. *Mutagenesis 18*(2), 159–166.

Dupuis, D. J. and D. C. Hamilton (2000). Regression residuals and test statistics: assessing naive outlier deletion. *The Canadian Journal of Statistics 28*(2), 259–275.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association 82*(397), 171–185.

Ejchart, A. and N. Sadlej-Sosnowska (2003). Statistical evaluation and comparison of comet assay results. *Mutation Research 534*, 85–92.

European Parliament (2010). Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes [2010] OJ L 276/33.

Fan, F. Y. (2013). *powerAnalysis: Power analysis in experimental design*. R package version 0.2.

Faul, F., E. Erdfelder, A. Buchner, and A.-G. Lang (2009). Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods 41*, 1149–1160.

Faul, F., E. Erdfelder, A.-G. Lang, and A. Buchner (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods 39*, 175–191.

Feinstein, A. R. and D. V. Cicchetti (1990). High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology 43*(6), 543–549.

Festing, M. F. W., P. Overend, R. G. Das, M. C. Borja, and M. Berdoy (2002). *The Design of Animal Experiments*. London, UK: Laboratory Animals Ltd.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin 76*(5), 378–382.

Food and Drug Administration (FDA) (2001). Guidance for industry: Bioanalytical method validation. http://www.fda.gov/cder/guidance/index.htm.

Fox, J. (2002). *An R and S-plus Companion to Applied Regression*. Thousand Oaks, CA: Sage.

Fox, J. and S. Weisberg (2011). *An R Companion to Applied Regression* (Second ed.). Thousand Oaks CA: Sage. R package version 1.1-1.

Galton, F. (1892). *Finger Prints*. London: Macmillan.

Gardenier, J. and D. Resnik (2002). The misuse of statistics: Concepts, tools, and a research agenda. *Accountability in Research: Policies and Quality Assurance 9*(2), 65–74.

Gaskin, C. J. and B. Happell (2013). Power of mental health nursing research: A statistical analysis of studies in the *International Journal of Mental Health Nursing*. *International Journal of Mental Health Nursing 22*, 69–75.

Giuffrida, M. A. (2014). Type II error and statistical power in reports of small animal clinical trials. *Journal of the American Medical Association 244*(9), 1075–1080.

Gray, J. B. and W. H. Woodall (1994). The maximum size of standardized and internally studentized resiudals in regression analysis. *The American Statistician 48*(2), 111–113.

Guérard, M., C. Marchand, and U. Plappert-Helbig (2014). Influence of experimental conditions on data variability in the liver comet assay. *Environmental and Molecular Mutagenesis 55*, 114–121.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology 61*, 29–48.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics 16*(3), 927–953.

Halvorsen, K. B. (2013). *asypow: Calculate Power Utilizing Asymptotic Likelihood Ratio Methods*. S original by Barry W. Brown and James Lovato and Kathy Russel. R package version 2013.9-1.

Hardin, J. W. and J. M. Hilbe (2007). *Generalized Linear Models and Extensions* (Second ed.). USA: Stata Press.

Harrell Jr, F. E. (2014). *rms: Regression Modeling Strategies*. R package version 4.2-0.

Hau, J. and S. J. Schapiro (2014). *Handbook of Laboratory Animal Science: Animal Models* (Third ed.), Volume 3. USA: CRC Press.

Hawkins, D., E. Gallacher, and M. Gammell (2013). Statistical power, effect size and animal welfare: recommendations for good practice. *Animal Welfare 22*, 339–344.

Hoaglin, D. C. and R. E. Welsch (1978). The hat matrix in regression and anova. *The American Statistician 32*(1), 17–22.

Hofmeister, E. H., J. King, M. R. Read, and S. C. Budsberg (2007). Sample size and statistical power in the small-animal analgesia literature. *Journal of Small Animal Practice 48*, 76–79.

Hubbard, R. and M. J. Bayarri (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician 57*(3), 171–178.

Jennions, M. D. and A. P. Møller (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology 14*(3), 438–445.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1994). *Continuous univariate distributions* (Second ed.), Volume 1 of *Wiley series in probability and mathematical statistics*. John Wiley & Sons.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1995). *Continuous univariate distributions* (Second ed.), Volume 2 of *Wiley series in probability and mathematical statistics*. John Wiley & Sons.

Johnson, R., J. Freund, and I. Miller (2010). *Probability and Statistics for Engineers* (eight ed.). Prentice Hall.

Julious, S. A. (2004). Tutorial in biostatistics: Sample sizes for clinical trials with normal data. *Statistics in Medicine 23*, 1921–1986.

Kenny, D. A. and C. M. Judd (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin 99*(3), 422–431.

Koletsi, D., N. Pandis, and P. S. Fleming (2014). Sample size in orthodontic randomized controlled trials: are numbers justified. *European Journal of Orthodontics 36*, 67–73.

Kotz, S., N. Balakrishnan, and N. L. Johnson (2000). *Continuous Multivariate Distributions, Models and Applications*. Wiley series in probability and statistics: Applied probability and statistics. John Wiley & Sons.

Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology*. Beverly Hills, CA: Sage.

Kromrey, J. D. and W. B. Dickinson (1996). Detecting unit of analysis problems in nested designs: statistical power and type I error rates of the F test for groups-within-treatments effects. *Educational and Psychological Measurement 56*(2), 215–231.

Kumaravel, T. S. and A. N. Jha (2006). Reliable Comet assay measurements for detecting dna damage induced by ionising radiation and chemicals. *Mutation Research 605*, 7–16.

Labes, D. (2014). *PowerTOST: Power and Sample size based on two one-sided t-tests (TOST) for (bio)equivalence studies*. R package version 1.1-11.

Lenth, R. (2009). Java Applets for Power and Sample Size [Computer software]. Retrieved May 29, 2014, from http://www.stat.uiowa.edu/~rlenth/Power.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician 55*(3), 187–193.

Lenth, R. V. (2007). Statistical power calculations. *Journal of Animal Science 85*, E24–E29.

Lin, L. (2007). Overview of agreement statistics for medical devices. *Journal of Biopharmaceutical Statistics 18*(1), 126–144.

Livingston, E. H. and L. Cassidy (2005). Statistical power and estimation of the number of required subjects for a study based on the *t*-test: A surgeon's primer. *Journal of Surgical Research 126*, 149–159.

Lovell, D. P. and T. Omori (2008). Statistical issues in the use of the comet assay. *Mutagenesis 23*(3), 171–182.

Lovell, D. P., G. Thomas, and R. Dubow (1999). Issues related to the experimental design and subsequent statistical analysis of in vivo and in vitro comet studies. *Teratogenesis, Carcinogenesis, and Mutagenesis 19*, 109–119.

Margolin, B. H. (1977). The distribution of internally studentized statistics via laplace transform inversion. *Biometrika 64*(3), 573–582.

Mayr, S., E. Erdfelder, A. Buchner, and F. Faul (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods 39*, 175–191.

Millard, S. P. (2013). *EnvStats: Package for Environmental Statistics, including US EPA Guidance*. R package version 1.0-2.

Montgomery, D. C. (2005). *Design and Analysis of Experiments* (sixth ed.). USA: John Wiley & Sons, Inc.

Mosteller, F. (1946). On some useful "inefficient" statistics. *Annals of Mathematical Statistics 17*(4), 377–408.

Musca, S. C., R. Kamiejski, A. Nugier, A. Méot, A. Er-Rafiy, and M. Brauer (2011). Data with hierarchical structure: impact of intraclass correlation and sample size on type-I error. *Frontiers in Psychology 2*(74).

Nawarathna, L. S. and P. K. Choudhary (2013). Measuring agreement in method comparison studies with heteroscedastic measurements. *Statistics in Medicine 32*, 5156–5171.

Neyman, J. and E. S. Pearson (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika 20A*(1/2), 175–240.

Neyman, J. and E. S. Pearson (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika 20A*(3/4), 263–294.

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature 506*(7487), 150–152.

Olive, P. L. and J. P. Banáth (2006). The comet assay: a method to measure dna damage in individual cells. *Nature Protocols 1*(1), 23–29.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing Series. New York, NY: Springer-Verlag.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications* (Second ed.). Wiley series in probability and mathematical statistics. USA: John Wiley & Sons.

Rizopoulos, D. and S. Tsonaka (2009). *grouped: Regression Analysis of Grouped and Coarse Data*. R package version 0.6-0.

Russell, W. M. S. and R. L. Burch (1959). *The Principles of Humane Experimental Technique*. London: Methuen.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika 6*(5), 309–316.

Schuster, C. (2005). Dispersion-weighted kappa: an integrative framework for metric and nominal scale agreement coefficients. *Psychometrika 70*(1), 135–146.

Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly 17*, 321–325.

Sena, E. S., H. B. van der Worp, P. M. W. Bath, D. W. Howells, and M. R. Macleod (2010). Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLOS Biology 8*(3).

Shapiro, S. S. and M. B. Wilk (1965). An analysis of variance test for normality (complete samples). *Biometrika 52*(3/4), 591–611.

Sim, J. and C. C. Wright (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy 85*(3), 257–268.

Smeeton, N. C. (1985). Early history of the kappa statistic. *Biometrics 41*, 795.

Smith, C. C., D. J. Adkins, E. A. Martin, and M. R. O'Donovan (2008). Recommendations for design of the rat comet assay. *Mutagenesis 23*(3), 233–240.

Smith, D. R., I. C. W. Hardy, and M. P. Gammell (2011). Power rangers: no improvement in the statistical power of analyses published in *Animal Behaviour*. *Animal Behaviour 81*(1), 347–352.

Strasak, A. M., Q. Zaman, K. P. Pfeiffer, G. Göbel, and H. Ulmer (2007). Statistical errors in medical research - a review of common pitfalls. *Swiss Medical Weekly 137*(3-4), 44–49.

The G*Power Team (2014). G*power 3.1 manual. http://www.gpower.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPowerManual.pdf. Retrieved May 29, 2014.

Tjur, T. (2009). Coefficients of determination in logistic regression models - a new proposal: The coefficient of discrimination. *The American Statistician 63*(4), 366–372.

Uebersax, J. S. (1992). Modeling approaches for the analysis of observer agreement. *Investigative radiology 27*(9), 738–743.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (Fourth ed.). New York: Springer.

Verde, P. E., L. A. Geracitano, L. L. Amado, C. E. Rosa, A. Bianchini, and J. M. Monserrat (2006). Application of public-domain statistical analysis software for evaluation and comparison of comet assay data. *Mutation Research 604*, 71–82.

Versi, E. (1992). "Gold standard" is an appropriate term. *British medical journal 305*(6846), 187–187.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika 29*(3/4), 350–362.

Wiklund, S. J. and E. Agurell (2003). Aspects of design and statistical analysis in the comet assay. *Mutagenesis 18*(2), 167–175.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association 22*(158), 209–212.

Yates, F. and M. J. R. Healy (1964). How should we reform the teaching of statistics. *Journal of the Royal Statistical Society. Series A 127*(2), 199–210.

Appendix A

Assessment of the type I error rate when ignoring the hierarchical structure of in vivo Comet assay data

Hansen, M. K. and Kulahci, M. (2014). Assessment of the type I error rate when ignoring the hierarchical structure of *in vivo* Comet assay data. In Peter Linde (ed.), *Symposium i Anvendt Statistik* [Symposium in Applied Statistics], University of Copenhagen, p. 83–92.

Assessment of the type I error rate when ignoring the hierarchical structure of *in vivo* Comet assay data

Merete K. Hansen and Murat Kulahci
DTU Compute, Technical University of Denmark

Abstract Damages to our DNA in terms of DNA strand breaks can be assessed with a technique known as the Comet assay. The experimental design of Comet assay studies are often hierarchically structured, however it does not seem to be commonly accounted for in the statistical analysis. Disregarding the hierarchical structure inflates the type I error rate considerably. Different combinations of the factors as they appear in a literature study result in type I error rates up to 0.51 and for all combinations the type I error rate is greater than the nominal α at 0.05. Closed-form expressions based on scaled F-distributions using the Welch-Satterthwaite approximation are derived to examine in which way the type I error rate is affected. These results are intended to motivate researchers to reconsider the analysis of hierarchical data when needed.

Introduction

Damage to our DNA occurs continuously due to both endogenous (e.g. metabolic processes) and exogenous (e.g. environmental agents) factors. DNA repair mechanisms are effective and constantly active, but some damages are irreparable. Accumulation of damages to the DNA may eventually become hazardous, as it may lead to unregulated cell division and tumors may evolve. The Comet assay is a rapid and sensitive technique for measuring DNA strand breaks within mammalian cells. The name of the assay originates from the images of comet-like structures that emerge due to DNA migration during electrophoresis of treated cells.

A common design of the *in vivo* Comet assay entails hierarchically structured data. However, this does not seem to be accounted for in the statistical analysis. This led us to investigate the implications in terms of the type I error rate when the hierarchical structure of the data is disregarded. The aim of this study was to to provide closed-form expressions for the type I error rate and to investigate whether the type I error rate considerably exceeded the nominal α. It is our hope that the results of this study can be used to motivate researchers to reconsider the statistical analysis when relevant. As similarly structured data appear in various research areas the results of this study may be equally relevant in other fields.

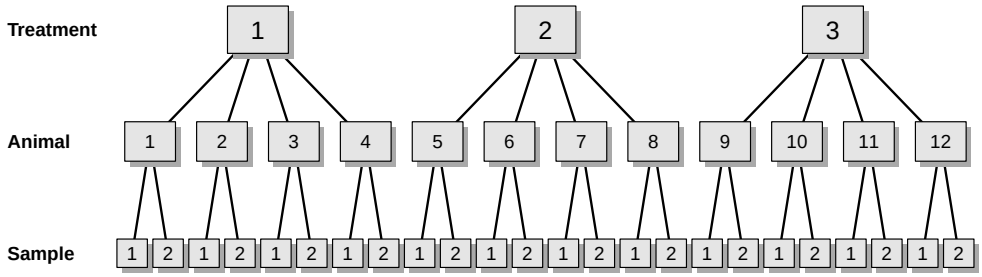

Figure 1: Outline of the design commonly used in Comet assay studies. This example shows three treatment groups, four animals per treatment and two samples per animal. For each sample a number of cells are scored, usually in the range of 50-100 cells.

Experimental design

A common design of *in vivo* Comet assay studies is illustrated in Figure 1. Animals are randomly assigned to one of a number of different treatment groups. These treatment groups often include one negative control group, one positive control group and dose groups where increasing doses of the compound of interest are administered to the animals. From each animal one or more samples are collected and from each sample a number of cells are scored.

Due to this setup the data have a hierarchical structure, that is, the sample of cells is nested within animal that in turn is nested within treatment. Often the interest lies in the assessment of the genotoxic effect potentially induced by the specific doses of the tested compound. The animals used in the study is not of particular interest but merely act as representatives of the general population of that species.

Literature study

To investigate how data are analyzed in practice a literature study was carried out. Papers were retrieved from the search engine Web of Science with *title: in vivo* and *topic: Comet assay*. Journal papers from January 2012 until October 2013 were considered. Papers conducting *in vivo* Comet assay studies with a experimental setup similar to Figure 1 were included.

Throughout the papers the execution of the experiment was well-described. This apply in particular to non-statistical aspects but also information about the number of treatment groups, number of animals in each group, number of samples and number of cells per sample were often clearly stated. Regarding the statistical analysis of the Comet assay data it was in general not easy to determine how it was conducted. Most often it was

briefly stated that data were analyzed with *one-way ANOVA* (46 %), *ANOVA* (31 %) or *Kruskal-Wallis test* (19 %). From the brief description provided in the papers we were not able to understand exactly how the analyses were performed. None of the papers defined a statistical model and no test statistics, degrees of freedom or other pointers were given. Some papers (65 %) indicated calculation of a summary measure such as the mean or median prior to the statistical analysis, but in most cases it was not clear how it was done, i.e. if the summary measure was calculated for each sample, for each animal etc. None of the papers mentioned mixed models, repeated measures ANOVA, random effects, nested effects or the like.

Modeling Comet assay data

One way to analyze data is to summarize the % tail DNA distribution for each sample into a single summary statistic and use this measure in a subsequent analysis. Due to the hierarchical structure of data and the randomly selected animals, a suitable analysis of the summarized data is a linear mixed-effects model with treatment as a fixed effect and animal as a random effect and with animal nested within treatment:

$$y_{ijk} = \mu + \tau_i + \beta_{(i)j} + \varepsilon_{(ij)k} \tag{1}$$

where

$$i = 1, ..., a,\ j = 1, ..., b,\ k = 1, ..., n,$$

$$\beta_{(i)j} \sim N(0, \sigma^2_\beta), \quad \varepsilon_{(ij)k} \sim N(0, \sigma^2).$$

y_{ijk} is the summary statistic of interest calculated for each sample and μ and τ_i are the fixed effects for the intercept and treatment, respectively. $\beta_{(i)j}$ is the random effect of the jth animal nested within the ith treatment and $\varepsilon_{(ij)k}$ is the within-group error. The parentheses in the subscripts indicate the nesting structure with the parent level(s) given inside the parentheses (Montgomery, 2005).

The literature study did not indicate that this model is commonly employed. Rather, it seems that data often are analyzed by means of a one-way ANOVA or the non-parametric equivalent Kruskal-Wallis test. In general, it was not clear if the raw scores (one for each cell), a summary measure calculated for each sample or a summary measure calculated for each animal was used in the analysis. In this study we will investigate the implications if data summarized for each sample are analyzed by means of a one-way ANOVA.

Notation

Let $Y_{ij.} = \sum_{k=1}^{n} y_{ijk}$, $Y_{i..} = \sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$, $Y_{...} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$ and let $\overline{Y}_{ij.} = \frac{1}{n}Y_{ij.}$, $\overline{Y}_{i..} = \frac{1}{bn}Y_{i..}$, $\overline{Y}_{...} = \frac{1}{abn}Y_{...}$. If $V \sim c\chi^2(\nu,\lambda)$ then V follows a scaled non-central χ^2-distribution with ν degrees of freedom, scaling parameter c and non-centrality parameter λ. If $c = 1$ and $\lambda = 0$ then we say that V follows a non-scaled central χ^2-distribution. If $W \sim cF(\nu_1,\nu_2,\lambda)$ then W has a scaled non-central F-distribution with ν_1 and ν_2 degrees of freedom, scaling parameter c and non-centrality parameter λ. The cumulative distribution function of W evaluated at w is denoted $G(w;\nu_1,\nu_2,\lambda)$ when W follows a non-scaled distribution or $G_s(w;\nu_1,\nu_2,\lambda)$ if the distribution is scaled. If $W \sim F(\nu_1,\nu_2)$ then W has a non-scaled central F-distribution with the critical value $F_{\alpha;\nu_1,\nu_2}$ being the $(1-\alpha)$th quantile such that $G(F_{\alpha;\nu_1,\nu_2};\nu_1,\nu_2) = 1-\alpha$.

Disregarding the hierarchical data structure

In Comet assay studies it is our impression that rather than model (1) the fixed-effects model is employed

$$y_{ij^*} = \mu + \tau_i + \varepsilon_{ij^*} \tag{2}$$

where $i = 1,...,a$, $j^* = 1,...,bn$ and $\varepsilon_{ij^*} \sim N(0,\sigma^{*2})$. This model typically underlies what is referred to as a one-way ANOVA. The F-statistic is calculated as

$$F_{\text{anova}} = \frac{bn\sum_{i=1}^{a}\left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2/(a-1)}{\sum_{i=1}^{a}\sum_{j^*=1}^{bn}\left(Y_{ij^*} - \overline{Y}_{i\cdot}\right)^2/(a(bn-1))} \tag{3}$$

which is expressed within the framework of model (1) as

$$F_{\text{anova}} = \frac{bn\sum_{i=1}^{a}\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right)^2/(a-1)}{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk} - \overline{Y}_{i\cdot\cdot}\right)^2/(a(bn-1))} \tag{4}$$

The denominator of (4) can be rewritten as

$$\left\{n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot}\right)^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk} - \overline{Y}_{ij\cdot}\right)^2\right\}/(a(b-1)+ab(n-1)) \tag{5}$$

implying that sum of squares and the degrees of freedom in the denominator is attributable both to the animal and the error part.

A nice feature of fitting data with model (1) is that the sum of squares in the numerator and denominator of the relevant F-statistic both follow χ^2-distributions that are scaled by $n\sigma_\beta^2+\sigma^2$, that is, they cancel out and the ratio follows a central non-scaled F-distribution. This is not the case for the F_{anova}-statistic in (4) as the sum of squares follow χ^2-distributions that are scaled differently. The sum of squares in the numerator is distributed as

$$bn\sum_{i=1}^{a}\left(\overline{Y}_{i\cdot\cdot}-\overline{Y}_{\cdot\cdot\cdot}\right)^2\sim\left(n\sigma_\beta^2+\sigma^2\right)\chi^2\left(a-1,\lambda\right), \tag{6}$$

where

$$\lambda=\frac{bn\sum\limits_{i=1}^{a}\tau_i^2}{n\sigma_\beta^2+\sigma^2}. \tag{7}$$

Since Y_{ijk} are not independent the sum of squares in the denominator of (4) does not follow the usual χ^2-distribution (see Box (1954) for details). However, looking separately at the two terms in the numerator of (5) gives

$$n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{Y}_{ij\cdot}-\overline{Y}_{i\cdot\cdot}\right)^2\sim\left(n\sigma_\beta^2+\sigma^2\right)\chi^2(a(b-1)), \tag{8}$$

and

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk}-\overline{Y}_{ij\cdot}\right)^2\sim\sigma^2\chi^2(ab(n-1)), \tag{9}$$

that is, $\sum\limits_{i=1}^{a}\sum\limits_{j=1}^{b}\sum\limits_{k=1}^{n}\left(Y_{ijk}-\overline{Y}_{i\cdot\cdot}\right)^2$ is a linear combination of independent χ^2-distributed random variables. An approximate distribution is obtained using the rationale of the Welch-Satterthwaite approximation (Welch, 1938; Satterthwaite, 1941; Box, 1954). The sum of squares is approximated by a scaled χ^2-distribution

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk}-\overline{Y}_{i\cdot\cdot}\right)^2\dot{\sim}\,c\chi^2(\nu) \tag{10}$$

and c and ν are found by matching the first two moments, so that

$$c=\frac{(b-1)(n\sigma_\beta^2+\sigma^2)^2+b(n-1)(\sigma^2)^2}{(b-1)(n\sigma_\beta^2+\sigma^2)+b(n-1)\sigma^2} \tag{11}$$

and

$$\nu=\frac{a((b-1)(n\sigma_\beta^2+\sigma^2)+b(n-1)\sigma^2)^2}{(b-1)(n\sigma_\beta^2+\sigma^2)^2+b(n-1)(\sigma^2)^2}, \tag{12}$$

where ν is known as the effective degrees of freedom (Satterthwaite, 1941). A ratio of χ^2-distributed random variables each divided by its degrees of freedom is F-distributed.

However, the sum of squares in the denominator of (4) is not divided by its effective degrees of freedom ν but by $a(bn-1)$, so that

$$F_{\text{anova}} = \frac{bn\sum_{i=1}^{a}\left(\overline{Y}_{i\cdot\cdot}-\overline{Y}_{\cdots}\right)^2/(a-1)}{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk}-\overline{Y}_{i\cdot\cdot}\right)^2/\nu} \cdot \frac{a(bn-1)}{\nu} \tag{13}$$

In addition, adjusting for the different scaling of the distributions of the numerator (scaled by $n\sigma_\beta^2+\sigma^2$) and denominator (scaled by c) gives an approximate distribution of F_{anova}

$$F_{\text{anova}} \dot{\sim} \frac{a(bn-1)}{\nu}\frac{n\sigma_\beta^2+\sigma^2}{c}F(a-1,\nu,\lambda), \tag{14}$$

and inserting ν and c gives

$$F_{\text{anova}} \dot{\sim} \xi\, F(a-1,\nu,\lambda), \tag{15}$$

where

$$\xi = \frac{(bn-1)(n\sigma_\beta^2+\sigma^2)}{(b-1)(n\sigma_\beta^2+\sigma^2)+b(n-1)\sigma^2}. \tag{16}$$

Since the expected value of an F random variable with ν_1 and ν_2 degrees of freedom is $E(F)=\frac{\nu_2(\nu_1+\lambda)}{\nu_1(\nu_2-2)}$ (Johnson et al., 1995), then

$$E(F_{\text{anova}}) \approx \xi\,\frac{\nu}{(a-1)(\nu-2)}\left(a-1+\frac{bn\sum_{i=1}^{a}\tau_i^2}{n\sigma_\beta^2+\sigma^2}\right) \tag{17}$$

and for sufficiently large ν

$$E(F_{\text{anova}}) \approx \xi\left(1+\frac{bn\sum_{i=1}^{a}\tau_i^2}{(a-1)(n\sigma_\beta^2+\sigma^2)}\right) \tag{18}$$

which under H_0 reduces to

$$E(F_{\text{anova}}) \approx \xi. \tag{19}$$

When $\sigma_\beta^2=0$ then $\xi=1$ and under H_0 then $E(F_{\text{anova}})\approx 1$. For $\sigma_\beta^2>0$ then $\xi>1$ implying that $E(F_{\text{anova}})>1$.

In practice, when data are analyzed with a one-way ANOVA the observed F_{anova}-statistic is (incorrectly) compared to a critical value obtained from an unscaled F-distribution, $F_{\alpha;a-1,a(bn-1)}$. The approximate type I error rate is found as

$$\text{Type I error rate} \approx 1-G_s(F_{\alpha;a-1,a(bn-1)};a-1,\nu), \tag{20}$$

where G_s refers to the scaled cumulative distribution function of F_{anova} given in (15) with $\lambda = 0$, since the type I error rate is defined under H_0.

Multiplying the scaled F-distribution by ξ^{-1} is a monotonic transformation, hence the type I error rate can also be calculated as

$$\text{Type I error rate} \approx 1 - G\left(\xi^{-1} F_{\alpha;a-1,a(bn-1)}; a-1, \nu\right) \tag{21}$$

and the type I error rate can be obtained by means of a non-scaled F-distribution, which is readily available in most statistical software.

The type I error rate can also be expressed in terms of the ratio of the variance components

$$\sigma^2_{\text{ratio}} = \frac{\sigma^2_\beta}{\sigma^2} \tag{22}$$

The effective degrees of freedom and the inverse scaling factor is then found as

$$\nu = \frac{a((b-1)(n\sigma^2_{\text{ratio}}+1)+b(n-1))^2}{(b-1)(n\sigma^2_{\text{ratio}}+1)^2+b(n-1)} \tag{23}$$

and

$$\xi = \frac{(bn-1)(n\sigma^2_{\text{ratio}}+1)}{(b-1)(n\sigma^2_{\text{ratio}}+1)+b(n-1)} \tag{24}$$

implying that the distribution of the F_{anova}-statistic and hence the type I error rate is influenced by the relative magnitudes of σ^2_β and σ^2.

In the special case where $n = 1$ (one sample per animal) then $\nu = a(b-1)$ and $\xi = 1$. In the case of $\sigma^2_\beta = 0$ then $\nu = a(bn-1)$ and $\xi = 1$. In both cases the approximate distribution in (15) becomes the usual (appropriate) F-distribution and the type I error rate in (21) becomes α.

Results

Table 1 summarizes the type I error rate for different combinations of treatment groups, animals per treatment, samples per animal and ratios of the variance components. The levels of the first three factors, i.e. treatment groups, animals and samples were selected among actual levels identified in the literature study, although not every combination of the three factors occurred. From an earlier study (Hansen et al., 2013) $\hat{\sigma}_{\text{ratio}} = 0.9 \approx 1$ and the levels of the ratio were selected as 0.5, 1 and 2 times this approximate estimate.

The assumed denominator degrees of freedom was calculated as $a(bn-1)$ as this is used when data are fitted model (2). The effective degrees of freedom ν, the scaling factor ξ and the approximate $E(F)$ was calculated from (23), (24) and (17), respectively.

Table 1: Type I error rate for different combinations of number of treatment groups (a), animals per treatment groups (b), samples per animal (n) and variance ratio $\sigma^2_{\text{ratio}} = \frac{\sigma^2_\beta}{\sigma^2}$. The effective denominator degrees of freedom ν, the scaling parameter ξ and the approximate $E(F)$ was calculated from (23), (24) and (17), respectively. The simulated type I error rate was based on 10000 simulations for each combination (each row) and the approximate type I error rate was found from (21). All approximate type I error rates were covered by the 95% confidence intervals for the simulated type I error rates except for two cases marked by asterisks. Type I error rates greater then 0.20 are marked in bold.

Treatment groups	Animals per treatment	Samples per animal	σ^2_{ratio}	Den DF $a(bn-1)$	Den DF ν	ξ	Approximate $E(F)$	Simulated type I error rate	Approximate type I error rate
2	4	2	0.5	14	12.50	1.40	1.67	0.094	0.094
2	4	2	1.0	14	10.90	1.61	1.98	0.118	0.120
2	4	2	2.0	14	9.14	1.84	2.36	0.150	0.148
2	4	3	0.5	22	17.96	1.77	2.00	0.139	0.137
2	4	3	1.0	22	14.29	2.20	2.56	0.186	0.183
2	4	3	2.0	22	10.85	2.65	3.25	**0.227**	**0.230**
2	8	2	0.5	30	26.89	1.36	1.47	0.090	0.092
2	8	2	1.0	30	23.69	1.55	1.70	0.120	0.114
2	8	2	2.0	30	20.21	1.74	1.94	0.140	0.138
2	8	3	0.5	46	37.56	1.72	1.81	0.131	0.133
2	8	3	1.0	46	30.25	2.09	2.24	0.178	0.174
2	8	3	2.0	46	23.54	2.48	2.71	**0.223**	**0.213***
6	4	2	0.5	42	37.50	1.40	1.48	0.147	0.149
6	4	2	1.0	42	32.71	1.61	1.72	**0.212**	**0.214**
6	4	2	2.0	42	27.42	1.84	1.99	**0.283**	**0.284**
6	4	3	0.5	66	53.89	1.77	1.84	**0.268**	**0.267**
6	4	3	1.0	66	42.86	2.20	2.31	**0.386**	**0.390**
6	4	3	2.0	66	32.55	2.65	2.83	**0.509**	**0.501**
6	8	2	0.5	90	80.67	1.36	1.40	0.149	0.144
6	8	2	1.0	90	71.07	1.55	1.60	0.194	**0.203***
6	8	2	2.0	90	60.62	1.74	1.80	**0.271**	**0.265**
6	8	3	0.5	138	112.69	1.72	1.75	**0.256**	**0.257**
6	8	3	1.0	138	90.75	2.09	2.14	**0.374**	**0.371**
6	8	3	2.0	138	70.61	2.48	2.55	**0.477**	**0.473**

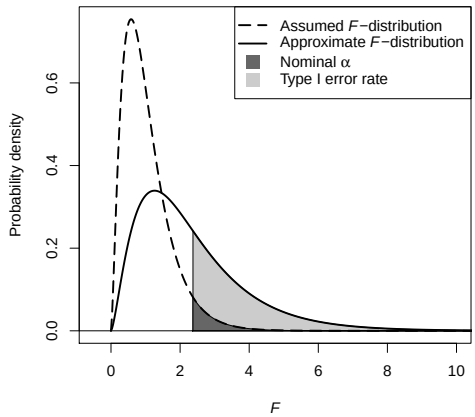

Figure 2: The F-distributions in case of 6 treatment groups, 4 animals per treatment, 3 samples per animal and $\sigma^2_{\text{ratio}} = 1$. The assumed F-distribution refers to the distribution from which the critical value is obtained. The approximate F-distribution is the distribution of F_{anova} as defined in (15). The approximate F-distribution has a heavier right tail implying that the type I error rate is greater than the nominal α at 0.05.

The simulated type I error rate was obtained by simulating data structured as in Figure 1, and for each combination (each row in Table 1) 10000 simulations were conducted. The approximate type I error rates were calculated from (21). Throughout the nominal α was 0.05.

In all cases the assumed denominator degrees of freedom were greater than the effective denominator degrees of freedom, ν, and furthermore $\xi > 1$. This imposes additional skewness to the F-distribution implying a heavier right tail as seen in Figure 2, which illustrates the F-distributions for 6 treatment groups, 4 animals, 3 samples and $\sigma^2_{\text{ratio}} = 1$.

Increasing the number of treatment groups enhanced the type I error rate considerably. The same was in evidence when the number of samples per animal were increased. Interestingly, the number of animals per treatment group did not affect the type I error rate noticeably. Increasing σ^2_{ratio} (increasing σ^2_β relative to σ^2) in general increased the type I error rate. All cases resulted in a type I error rate greater than the nominal α at 0.05. Most combinations gave type I error rates greater than 0.10 and almost half resulted in type I error rates greater than 0.20.

The validity of the approximate type I error rates was assessed by making an informal comparison to the simulated type I error rates. To quantify the simulation uncertainty the standard errors were calculated as $\text{se}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$ and Wald based 95% confidence intervals (CI) were obtained (not shown). The simulated type I error rates were between 0.090 and 0.509, hence the standard errors were between 0.003 and 0.005. In all but two cases the approximate type I error rates were covered by the CI for the simu-

lated type I error rates. This agrees with the expectation of 1 to 2 values falling outside the CI given the number of comparisons and the confidence level. The two cases not covered by the CI are marked with asterisks in Table 1. A 99% CI for the simulated type I error rates covered all the approximate type I error rates.

Discussion

This study aimed at addressing potential issues concerning the analysis of Comet assay data. First, from the literature study it was not possible to deduce exactly how data were analyzed, which impedes reproducibility and blurs the interpretation of the reported results. Even if some researchers analyze data properly, we find it likely that others (e.g. new researchers in the field) may be inspired by the insufficient description of the statistical modeling in the papers and thereby may fail to allow properly for the nested structure. Second, as we suspect that the nested structure in data is not accounted for in the statistical model we investigated the implications in terms of the type I error rate. Approximate formulas were derived to examine in which way the type I error rate was affected. Type I error rates for different combinations of the factors as they appeared in the literature study demonstrated that the inflation is in fact non-trivial and resulted in type I error rates up to 0.51.

Our objective was to illustrate the implications in a simple manner with the hope of motivating researchers within the field to reconsider the statistical modeling. As the design considered here is widespread across various scientific areas we believe that the results may be equally relevant to researchers in other fields.

References

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics 25*(2), 290–302.

Hansen, M. K., A. K. Sharma, M. Dybdahl, J. Boberg, and M. Kulahci (2013). *In vivo* comet assay - statistical analysis and power calculations of mice testicular cells. Submitted.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1995). *Continuous univariate distributions* (second ed.), Volume 2 of *Wiley series in probability and mathematical statistics*. John Wiley & Sons.

Montgomery, D. C. (2005). *Design and Analysis of Experiments* (sixth ed.). USA: John Wiley & Sons, Inc.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika 6*(5), 309–316.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika 29*(3/4), 350–362.

Appendix B

The type I error rate for Comet assay data when the hierarchical structure is disregarded

---

Hansen, M. K. and Kulahci, M. (2014). *The type I error rate for in vivo Comet assay data when the hierarchical structure is disregarded.* DTU Compute Technical Report No. 9. Department of Applied Mathematics and Computer Science, Technical University of Denmark.

The type I error rate for *in vivo* Comet assay data when the hierarchical structure is disregarded

DTU Compute Technical Report-2014-09

Merete Kjær Hansen* and Murat Kulahci

Department of Applied Mathematics and Computer Science, Technical University of Denmark, DK–2800 Kgs. Lyngby, Denmark

June 15, 2014

*Corresponding author, e-mail: mkha@dtu.dk

Abstract

The Comet assay is a sensitive technique for detection of DNA strand breaks. The experimental design of *in vivo* Comet assay studies are often hierarchically structured, which should be reflected in the statistical analysis. However, the hierarchical structure sometimes seems to be disregarded, and this imposes considerable impact on the type I error rate. This study aims to demonstrate the implications that result from disregarding the hierarchical structure. Different combinations of the factor levels as they appear in a literature study give type I error rates up to 0.51 and for all combinations the type I error rate is greater than the nominal α at 0.05. Closed-form expressions based on scaled F-distributions using the Welch-Satterthwaite approximation are provided to show how the type I error rate is affected. With this study we hope to motivate researchers to be more precise regarding the exposition of the statistical methodology and to suitably account for the hierarchical structure of Comet assay data whenever present.

Contents

1 Introduction

Damage to our DNA occurs continuously due to both endogenous (e.g. metabolic processes) and exogenous (e.g. environmental agents) factors. DNA repair mechanisms are effective and constantly active, but some damages are irreparable. Accumulation of damages to the DNA may eventually become hazardous, as it may lead to unregulated cell division and tumors may evolve (Jeggo and Löbrich, 2007). The Comet assay is a rapid and sensitive technique for measuring DNA strand breaks within mammalian cells. The name of the assay originates from the images of comet-like structures that emerge due to DNA migration during electrophoresis of treated cells (Kumaravel and Jha, 2006; Hartmann et al., 2003).

A common design of the *in vivo* Comet assay entails hierarchically structured data. However, this does not seem to be accounted for in the statistical analysis. This led us to investigate the implications in terms of the type I error rate when the hierarchical structure of the data is disregarded. The type I error rate for two different hierarchical structures were assessed and it was investigated whether the the type I error rate considerably exceeded the nominal α. Closed-form expressions are provided for one of these cases.

A literature study revealed that it was not possible to determine exactly how data were analyzed due to an inadequate description. This is unfortunate since it impedes reproducibility and blurs the interpretation of the reported results. Although some researchers may analyze data properly, we find it likely that others are inspired by the insufficient description in the papers and thereby unintentionally may fail to allow properly for the nested structure.

The aim of this study is twofold. First, we aim to shed light on the insufficient description of the statistical modeling that currently characterizes some papers describing Comet assay data. Second, the implications of disregarding the hierarchical structure of data in the statistical modeling are assessed.

All results and derivations in this report assume balanced data, i.e. that the number of observations in each subgroup are the same. This is usually endavoured in Comet assay studies and it is not uncommon for designed experiments in general.

The structure of the report is as follows. Section 2 describes a common design of Comet assay studies and the resulting inherent hierarchical nature of the collected data. Section 3 presents possible statistical models for fitting raw or summarized

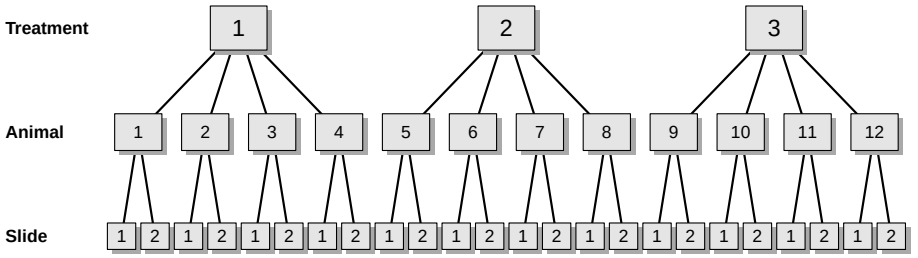

Figure 1: Outline of the design commonly used in Comet assay studies. This example shows three treatment groups, four animals per treatment and two slides per animal. For each slide a number of cells are scored, usually in the range of 50-100 cells.

Comet assay data. Section 4 describes a literature review examining the statistical analysis conducted in these published studies. Section 5 exposits the notation and relevant existing results that are used in this report. In Section 6 we look at the sampling distribution when a nested mixed-effects model is used to fit data. Section 7 provides simulated type I error rates when the hierarchical data structure is ignored in case of two different hierarchical data structures. Furthermore, closed-form expressions for the type I error rate for one of these cases is derived. Section 8 contains a discussion of the results. Some intermediate derivations are given in the Appendix.

2 Comet assay data

A common design of *in vivo* Comet assay studies is illustrated in Figure 1. Animals are randomly assigned to one of a number of different treatment groups. These treatment groups often include one negative control group, one positive control group and dose groups where increasing doses of the compound of interest are administered to the animals. For each animal there are a number of slides, in practice usually one to three slides, and from each slide a number of cells are scored.

This setup imposes a hierarchical structure of data, that is, the cells are nested within slide, that in turn is nested within animal, which again is nested within treatment. Often the interest lies in the assessment of the genotoxic effect potentially induced by the specific doses of the specific compound tested. The specific

animals used in the study is not of particular interest but merely act as representatives of the general population of that species.

50-100 cells are usually scored for each slide and the shape of the individual electrophoresed cells are very distinct. Cells can be scored both manually and automatically. One example of manual scoring is to categorise each cell in one of five categories ranging from 0 to 4 according to the shape of the cell, and a total sum is calculated for each slide or animal (Zan et al., 2013; Pesarini et al., 2013; Malta et al., 2012). Automated scoring is performed by imaging software. Popular end points are % tail DNA (percent DNA located in the Comet "tail") and the Olive tail moment, which is the product of the tail length and % tail DNA (Olive et al., 1990; Lovell and Omori, 2008). Most of the findings in the current report will be equally relevant for all types of end points assuming that they are normally distributed, possibly by transformation.

Sometimes, a summary statistic is calculated and used as response in the statistical modeling. A natural question that arises is which summary statistic to employ. Different summary statistics have been proposed, including the mean (Bright et al., 2011; Lovell et al., 1999; Wiklund and Agurell, 2003), the median (Bright et al., 2011; Lovell et al., 1999; Wiklund and Agurell, 2003; Duez et al., 2003), the 75th percentile (Lovell et al., 1999; Duez et al., 2003) and the 90th percentile (Lovell et al., 1999; Wiklund and Agurell, 2003). Also, to comply with the skewness of the within-slide distributions it has been suggested to log-transform the raw data prior to the summary calculations (Lovell and Omori, 2008). Although a few studies specifically address these issues, there is currently no consensus as to which statistic most appropriately summarizes data.

3 Statistical analysis of Comet assay data

Comet assay data can be analyzed in different ways. For some end points (e.g. % tail DNA and tail moment) data are heavily skewed and it has been suggested to model the data by means of the Weibull distribution (Ejchart and Sadlej-Sosnowska, 2003; Verde et al., 2006). In practice, it seems that only statistical methods relying on the normal distribution are used and three related statistical models valid for fitting Comet assay data are presented in the following. When data are balanced and normally distributed all three methods are equivalent. However, this requires that the statistical model matches data, i.e. if a summary statistic some-

how is calculated from the raw data this should appropriately be reflected in the model. Due to the assumption of normally distributed data it may be requisite to transform data prior to the statistical modeling.

3.1 Using raw cell scores as the response

When the raw cell scores are used as the response the hierarchical structure of data and the randomly selected animals should be properly accounted for. This can be done by employing a linear mixed-effects model with treatment as a fixed effect and animal and slide as random effects. Animal is nested within treatment and slide is nested within animal:

$$Y_{ijkl} = \mu + \tau_i + \beta_{(i)j} + \gamma_{(ij)k} + \varepsilon_{(ijk)l} \tag{1}$$

where

$$i = 1, ..., a,\ j = 1, ..., b,\ k = 1, ..., c,\ l = 1, ..., n,$$
$$\beta_{(i)j} \sim N(0, \sigma^2_\beta), \quad \gamma_{(ij)k} \sim N(0, \sigma^2_\gamma), \quad \varepsilon_{(ijk)l} \sim N(0, \sigma^2).$$

Y_{ijkl} is the $ijkl$th observation (one score for each cell) and μ and τ_i are the fixed effects for the intercept and treatment, respectively. $\beta_{(i)j}$ is the random effect of the jth animal nested within the ith treatment, $\gamma_{(ij)k}$ is the random effect of the kth slide nested within the ith treatment and jth animal and $\varepsilon_{(ij)k}$ is the within-group error. The parentheses in the subscripts indicate the nesting structure with the parent level(s) given inside the parentheses. See Montgomery (2005) for a more elaborate exposition of the linear mixed-effects model with nested effects.

3.2 Summarizing the response for each slide

Another way to analyze data is to summarize the % tail DNA distribution for each slide into a single summary statistic and use this measure in the subsequent analysis. Due to the hierarchical structure of data and the randomly selected animals a suitable analysis of the summarized data is a linear mixed-effects model with treatment as a fixed effect and animal as a random effect and with animal nested within treatment:

$$Y_{ijk} = \mu + \tau_i + \beta_{(i)j} + \varepsilon_{(ij)k} \tag{2}$$

where

$$i = 1, ..., a,\ j = 1, ..., b,\ k = 1, ..., n,$$
$$\beta_{(i)j} \sim N(0, \sigma^2_\beta), \quad \varepsilon_{(ij)k} \sim N(0, \sigma^2).$$

Y_{ijk} is the summary statistic of interest calculated for each slide and μ and τ_i are the fixed effects for the intercept and treatment, respectively. $\beta_{(i)j}$ is the random effect of the jth animal nested within the ith treatment and $\varepsilon_{(ij)k}$ is the within-group error.

3.3 Summarizing the response for each animal

A third option is to calculate a summary statistic for each animal and use this as the response. A suitable model is the fixed-effects model with treatment as a fixed effect:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \tag{3}$$

where

$$i = 1, ..., a,\ j = 1, ..., n,$$
$$\varepsilon_{ij} \sim N(0, \sigma^2).$$

Y_{ij} is the summary statistic of interest calculated for each animal, μ and τ_i are fixed effects for the intercept and treatment, respectively, and ε_{ij} is the within-group error.

4 Literature study

To investigate how data are analyzed in practice a literature study was carried out. Papers were retrieved from the search engine Web of Science with *title: in vivo* and *topic: Comet assay* from January 2012 until December 2013, which resulted in 95 papers. Of these, 47 papers conducted *in vivo* Comet assay studies with an experimental setup similar to Figure 1, and these were included in the current literature study.

Throughout the papers the execution of the experiment was well-described. This apply in particular to non-statistical aspects but also information about the number of treatment groups, number of animals per group, number of slides per animal and number of cells per slide were often clearly stated.

Regarding the statistical analysis of the Comet assay data it was in general not easy to determine how it was conducted. None of the papers defined a statistical model and no test statistics, degrees of freedom or other pointers were given. Most often it was briefly stated that data were analyzed with *one-way ANOVA* (45%), *ANOVA* (21%) or *Kruskal-Wallis test* (15%). The remaining papers predominantly used Student's t-test (also in case of more than two treatments), *Mann-Whitneys U test* or *post-hoc* tests such as Dunnett's test without preceeding use of other statistical models. None of the papers mentioned mixed models, repeated measures ANOVA, random effects, nested effects or the like.

18 papers (38%) stated "Results are expressed as mean $\pm$ SD" (or mean $\pm$ SE) or something similarly phrased. However, it was not clear how it was calculated, i.e. if these measures were calculated for each slide, for each animal etc. Also, it was not clear whether the statement was related to the tables presenting data or the statistical analysis of data. In some of these cases other summary statistics were calculated prior to the statistical analysis, i.e. in at least some cases it seems only to concern the tables summarizing data.

23 papers (49%) calculated a summary measure prior to the statistical analysis. Of these, only 15 papers (65%) clearly stated how it was done, and in these cases a summary statistic most often was calculated for each animal; that amounts to 32% of all papers that were included in the literature study. In the other 8 papers (35%) it was not possible to deduce how the summary statistic was calculated, i.e. if it was calculated per slide, per animal etc.

In 24 papers (51%) it seemed as no summary statistic was calculated prior to the statistical analysis.

The imprecise description of the statistical analysis in these papers is of a concern to us for two reasons. First, indistinctness of the methodology impedes both reproducibility and a proper interpretation of the results. Second, the combination of the lack of a calculated summary statistic and the reported statistical models that are used strongly indicates that the hierarchical structure is not properly accounted for in the statistical analysis. Although some researchers may analyze data properly, we find it likely that others are inspired by the inadequate description that implicitly suggests not to account for the hierarchical structure of data.

We performed this study to accommodate these exact concerns. By bringing these issues into focus we hope to motivate researchers to elaborate the description of

the statistical methodology. Furthermore, we wish to create awareness of the implications of ignoring potential hierarchical structure of data.

5 Notation and existing results

If $V \sim c\chi^2(\nu, \lambda)$ then V is said to follow a scaled non-central χ^2-distribution with ν degrees of freedom, scaling parameter c and non-centrality parameter λ. If $c = 1$ and $\lambda = 0$ then we say that V follows a non-scaled central χ^2-distribution. If $W \sim cF(\nu_1, \nu_2, \lambda)$ then W has a scaled non-central F-distribution with ν_1 and ν_2 degrees of freedom, scaling parameter c and non-centrality parameter λ. The cumulative distribution function of W evaluated at w is denoted $G(w; \nu_1, \nu_2, \lambda)$ when W follows a non-scaled distribution or $G_s(w; \nu_1, \nu_2, \lambda)$ if the distribution is scaled. If $W \sim F(\nu_1, \nu_2)$ then W has a non-scaled central F-distribution with the critical value $F_{\alpha;\nu_1,\nu_2}$ being the $(1-\alpha)$th quantile such that $G(F_{\alpha;\nu_1,\nu_2}; \nu_1, \nu_2) = 1 - \alpha$.

Let $X_1, X_2, ..., X_n$ be independent random variables normally distributed with expected values $E(X_1), E(X_2), \ldots, E(X_n)$ and common variance $\mathsf{Var}(X_1) = \mathsf{Var}(X_2) = \cdots = \mathsf{Var}(X_n) = \mathsf{Var}(X)$. Also, let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $E(\overline{X}) = \overline{E(X)} = \frac{1}{n}\sum_{i=1}^{n} E(X_i)$. Then

$$V = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\mathsf{Var}(X)} \sim \chi^2(n-1, \lambda), \tag{4}$$

where λ is the non-centrality parameter given as

$$\lambda = \frac{\sum_{i=1}^{n} \left(E(X_i) - E(\overline{X})\right)^2}{\mathsf{Var}(X)}$$

(Johnson et al., 1995). Furthermore, if $V_1 \sim \chi^2(\nu_1, \lambda_1)$ and $V_2 \sim \chi^2(\nu_2, \lambda_2)$ are independent random variables, then according to the reproductive property of the χ^2-distribution the sum is distributed as

$$V_1 + V_2 \sim \chi^2(\nu_1 + \nu_2, \lambda_1 + \lambda_2) \tag{5}$$

(Johnson et al., 1995; Dobson, 2002). The ratio of two independent χ^2-distributed random variables, $V_1 \sim \chi^2(\nu_1, \lambda)$ and $V_2 \sim \chi^2(\nu_2)$, each divided by its degrees of freedom follows an F-distribution with ν_1 and ν_2 degrees of freedom

$$W = \frac{V_1/\nu_1}{V_2/\nu_2} \sim F(\nu_1, \nu_2, \lambda). \tag{6}$$

with the expected value

$$E(W) = \frac{\nu_2(\nu_1 + \lambda)}{\nu_1(\nu_2 - 2)} \tag{7}$$

(Johnson et al., 1995).

Now, let $y_{ij\cdot} = \sum_{k=1}^{n} y_{ijk}$, $y_{i\cdot\cdot} = \sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$, $y_{\cdots} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}$ and let $\bar{y}_{ij\cdot} = \frac{1}{n} y_{ij\cdot}$, $\bar{y}_{i\cdot\cdot} = \frac{1}{bn} y_{i\cdot\cdot}$, $\bar{y}_{\cdots} = \frac{1}{abn} y_{\cdots}$. The observations y_{ijk} and the group averages $\bar{y}_{ij\cdot}$, $\bar{y}_{i\cdot\cdot}$ and $\bar{y}_{\cdots}$ are realizations of the random variables Y_{ijk}, $\overline{Y}_{ij\cdot}$, $\overline{Y}_{i\cdot\cdot}$ and $\overline{Y}_{\cdots}$, respectively. They are distributed as

$$\begin{aligned}
Y_{ijk} &\sim N\left(\mu + \tau_i, \sigma_\beta^2 + \sigma^2\right) \\
\overline{Y}_{ij\cdot} &\sim N\left(\mu + \tau_i, \frac{n\sigma_\beta^2 + \sigma^2}{n}\right) \\
\overline{Y}_{i\cdot\cdot} &\sim N\left(\mu + \tau_i, \frac{n\sigma_\beta^2 + \sigma^2}{bn}\right) \\
\overline{Y}_{\cdots} &\sim N\left(\mu, \frac{n\sigma_\beta^2 + \sigma^2}{abn}\right)
\end{aligned} \tag{8}$$

See appendix A for details. Furthermore,

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \overline{Y}_{ij\cdot})^2 \sim \sigma^2\chi^2\left(ab(n-1)\right) \tag{9}$$

$$n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot})^2 \sim (\sigma_\beta^2 + \sigma^2)\chi^2\left(a(b-1)\right) \tag{10}$$

$$bn\sum_{i=1}^{a}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2 \sim (\sigma_\beta^2 + \sigma^2)\chi^2\left(a-1, \lambda\right) \tag{11}$$

where

$$\lambda = \frac{bn \sum_{i=1}^{a} \tau_i}{n\sigma_\beta^2 + \sigma^2}$$

See appendix B for details.

6 Hierarchical models for hierarchical data

In this section we will look into the behaviour of the sampling distribution when Comet assay data summarized for each slide (i.e. as described in section 3.2) are fitted a linear mixed-effects model as defined in (2).

The hypothesis of interest is concerning equality of the different dose groups

$$\begin{aligned} H_0&: \ \tau_1 = \tau_2 = \cdots = \tau_a = 0 \\ H_1&: \ \text{at least one } \tau_i \neq 0. \end{aligned}$$

We first consider the sum of squares attributable to the treatment effect. According to (11) then

$$\frac{bn \sum_{i=1}^{a} \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right)^2}{n\sigma_\beta^2 + \sigma^2} \sim \chi^2(a-1, \lambda), \tag{12}$$

where

$$\lambda = \frac{bn \sum_{i=1}^{a} \tau_i^2}{n\sigma_\beta^2 + \sigma^2}.$$

Considering the sum of squares reflecting the error component then according to (10)

$$\frac{n \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot}\right)^2}{n\sigma_\beta^2 + \sigma^2} \sim \chi^2(a(b-1)). \tag{13}$$

As stated in (6) a ratio of two independent χ^2-distributed random variables each divided by their corresponding degrees of freedom follows an F-distribution. It

can be shown with Fisher-Cochran's theorem (Rao, 1973) that (12) and (13) are independent, hence

$$F_{\text{mixed}} = \frac{\left\{ bn \sum_{i=1}^{a} \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot} \right)^2 / \left(n\sigma_\beta^2 + \sigma^2 \right) \right\} / (a-1)}{\left\{ n \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} \right)^2 / \left(n\sigma_\beta^2 + \sigma^2 \right) \right\} / (a(b-1))}$$

$$= \frac{bn \sum_{i=1}^{a} \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot} \right)^2 / (a-1)}{n \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} \right)^2 / (a(b-1))} \sim F\left(a-1, a(b-1), \lambda\right), \tag{14}$$

where

$$\lambda = \frac{bn \sum_{i=1}^{a} \tau_i^2}{n\sigma_\beta^2 + \sigma^2}. \tag{15}$$

According to (7) the expected value of (14) is

$$E(F_{\text{mixed}}) = \frac{a(b-1)}{(a-1)(a(b-1)-2)} \left(a - 1 + \frac{bn \sum_{i=1}^{a} \tau_i^2}{n\sigma_\beta^2 + \sigma^2} \right) \tag{16}$$

and for sufficiently large a or b then

$$E(F_{\text{mixed}}) \approx 1 + \frac{bn \sum_{i=1}^{a} \tau_i^2}{(a-1)(n\sigma_\beta^2 + \sigma^2)} \tag{17}$$

which under H_0 reduces to

$$E(F_{\text{mixed}}) \approx 1 \tag{18}$$

7 The type I error rate - Disregarding the hierarchical structure

A type I error occurs if H_0 is rejected when it indeed is true. A type II error occurs if H_0 is not rejected although it is false (i.e. H_1 is true). A type I error

is often considered the more serious of the two and is therefore controlled more strictly. The probability of making a type I error is also called the significance level and is denoted α (Johnson et al., 2010; Hogg et al., 2005).

If some of the model assumptions are violated the actual probability of making a type I error will differ from the pre-specified significance level. Therefore, we distinguish between the former, which also is called the actual α, and the latter, which is denoted the nominal α.

From our literature study it appears as data most often are analyzed with a one-way ANOVA or Kruskal-Wallis test. However, in many cases it also seems that a suiting summary measure is not used as the response. This combination violates the assumption of independence since the observations obtained from the same animal in that case will be correlated. In the following we will investigate the implications when a one-way ANOVA is used in the analysis of hierarchically structured Comet assay data, that is, when the response is the raw cell scores as described in section 3.1 or when the response is a summary measure for each slide as described in section 3.2. The type I error rate is obtained by simulation in case of raw cell scores. Closed-form expressions for the type I error rate are provided when the response is a summary measure for each slide. The type I error rates are calculated from these expressions and are validated by simulations.

7.1 Using raw cell scores as the response

Type I error rates are in the following obtained by simulating data with a structure as depicted in Figure 1. The simulated data are subsequently analyzed by means of a one-way ANOVA, i.e. data are fitted model (3).

Table 1 shows the type I error rates for different combinations of number of treatment groups, number of animals per treatment, number of slides per animals and number of cells per animal. The levels reflect the numbers that appeared in the literature study although not all exact combinations occured. The variance components used in the simulation study were $\sigma^2_\beta = 0.08$ (animal-to-animal variation), $\sigma^2_\gamma = 0.04$ (slide-to-slide variation) and $\sigma^2 = 2.92$. These variance component equals the estimates obtained by fitting model (1) to Comet assay data obtained from an earlier study (Hansen et al., 2014). The study used % tail DNA as end point and these estimates may thus not apply to data using other end points such as the Olive tail moment or tail length. Nonetheless, the results given here can be

Table 1: Type I error rate for different combinations of number of treatment groups, animals per treatment groups, slides per animal and cells per slide. The simulated type I error rate was based on 10000 simulations for each combination (each row). The variance components used in the simulations were $\sigma^2_\beta = 0.08$ (animal-to-animal variation), $\sigma^2_\gamma = 0.04$ (slide-to-slide variation) and $\sigma^2 = 2.92$

Treatment groups	Animals per treatment	Slides per animal	Cells per slide	Simulated type I error rate
2	4	2	50	0.335
2	4	2	100	0.474
2	4	3	50	0.397
2	4	3	100	0.535
2	8	2	50	0.330
2	8	2	100	0.464
2	8	3	50	0.398
2	8	3	100	0.532
6	4	2	50	0.747
6	4	2	100	0.909
6	4	3	50	0.840
6	4	3	100	0.950
6	8	2	50	0.758
6	8	2	100	0.905
6	8	3	50	0.838
6	8	3	100	0.950

used to give an impression of the implications when the hierarchical structure is disregarded.

As seen in table 1 the type I error rate is severely inflated in all cases. The lowest type I error rate for the combinations shown here occurs when we have the lowest number of observations, namely when there is two treatment groups, four animals per treatment, two slides per animal and 50 cells per slide. Increasing the number of animals per treatment group did not affect the type I error rate much. Increasing the number of treatment groups, number of slides per animal and number of cells per slide generally resulted in increasing type I error rates. The type I error rates are between 0.335 and 0.950 and all type I error rates are thus seriously inflated. In the best case a false positive is obtained more than 3 out of 10 times whereas in the most severe case a false positive occurs more than 9 out of 10 times.

7.2 Summarizing the response for each slide

We will in the following assess the type I error rate when a summary statistic is calculated for each slide and subsequently used as the response when model (3) is fitted. First, approximate closed-form expressions are derived which aid in disclosing how the different factors affect the type I error rate. Subsequently, approximate type I error rates for different combinations of the relevant factors are calculated from the closed-form expressions and shown together with simulated type I error rates.

7.2.1 Closed-form expressions for the type I error rate

Assume that a summary measure is calculated for each slide and the fixed-effects model is employed

$$Y_{ij^*} = \mu + \tau_i + \varepsilon_{ij^*} \tag{19}$$

where $i = 1, ..., a,\ \ j^* = 1, ..., bn$ and $\varepsilon_{ij^*} \sim N(0, \sigma^{*2})$. This model typically underlies what is referred to as a one-way ANOVA. The F-statistic is calculated as

$$F_{\text{anova}} = \frac{bn \sum_{i=1}^{a} \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2 / (a-1)}{\sum_{i=1}^{a} \sum_{j^*=1}^{bn} \left(Y_{ij^*} - \overline{Y}_{i\cdot}\right)^2 / (a(bn-1))} \tag{20}$$

which is expressed within the framework of model (2) as

$$F_{\text{anova}} = \frac{bn \sum_{i=1}^{a} \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot}\right)^2 / (a-1)}{\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} \left(Y_{ijk} - \overline{Y}_{i\cdot\cdot}\right)^2 / (a(bn-1))} \tag{21}$$

The denominator of (21) can be rewritten as

$$\left\{ n \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot}\right)^2 + \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} \left(Y_{ijk} - \overline{Y}_{ij\cdot}\right)^2 \right\} / \left(a(b-1) + ab(n-1)\right) \tag{22}$$

implying that sum of squares and the degrees of freedom in the denominator is attributable both to the animal and the error part.

A nice feature of the F_{mixed}-statistic given in (14) is that the sum of squares in the numerator and denominator both follow χ^2-distributions that are scaled by $n\sigma_\beta^2 + \sigma^2$, that is, they cancel out and the ratio follows a standard F-distribution. This is not the case for F_{anova}-statistic in (21) as the sum of squares follow χ^2-distributions that are scaled differently. The sum of squares in the numerator is distributed as

$$bn \sum_{i=1}^{a} \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot}\right)^2 \sim \left(n\sigma_\beta^2 + \sigma^2\right) \chi^2 \left(a-1, \lambda\right), \tag{23}$$

where λ is given in (15). Since Y_{ijk} are not independent the denominator of (21) does not follow the usual χ^2-distribution (see Box (1954) for details). However, looking separately at the two terms in the numerator of (22) gives

$$n \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot}\right)^2 \sim \left(n\sigma_\beta^2 + \sigma^2\right) \chi^2(a(b-1)), \tag{24}$$

and

$$\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} \left(Y_{ijk} - \overline{Y}_{ij\cdot}\right)^2 \sim \sigma^2 \chi^2(ab(n-1)), \tag{25}$$

that is, $\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} \left(Y_{ijk} - \overline{Y}_{i\cdot\cdot}\right)^2$ is a linear combination of independent χ^2-distributed random variables. An approximate distribution is obtained using the rationale of the Welch-Satterthwaite approximation (Welch, 1938; Satterthwaite, 1941; Box, 1954). The sum of squares is approximated by a scaled χ^2-distribution

$$\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} \left(Y_{ijk} - \overline{Y}_{i\cdot\cdot}\right)^2 \dot{\sim} c\chi^2(\nu) \tag{26}$$

where c and ν are found by matching the first two moments (see appendix C). Thus,

$$\begin{aligned} c &= \frac{a(b-1)(n\sigma_\beta^2 + \sigma^2)^2 + ab(n-1)(\sigma^2)^2}{a(b-1)(n\sigma_\beta^2 + \sigma^2) + ab(n-1)\sigma^2} \\ &= \frac{(b-1)(n\sigma_\beta^2 + \sigma^2)^2 + b(n-1)(\sigma^2)^2}{(b-1)(n\sigma_\beta^2 + \sigma^2) + b(n-1)\sigma^2} \end{aligned} \tag{27}$$

and

$$\nu = \frac{(a(b-1)(n\sigma_\beta^2+\sigma^2)+ab(n-1)\sigma^2)^2}{a(b-1)(n\sigma_\beta^2+\sigma^2)^2+ab(n-1)(\sigma^2)^2}$$

$$= \frac{a((b-1)(n\sigma_\beta^2+\sigma^2)+b(n-1)\sigma^2)^2}{(b-1)(n\sigma_\beta^2+\sigma^2)^2+b(n-1)(\sigma^2)^2} \tag{28}$$

where ν is known as the effective degrees of freedom (Satterthwaite, 1941). As previously mentioned a ratio of χ^2-distributed random variables each divided by its degrees of freedom are F-distributed. However, the sum of squares in the denominator of (21) is not divided by its effective degrees of freedom ν but by $a(bn-1)$, so that

$$F_{\text{anova}} = \frac{bn\sum_{i=1}^{a}\left(\overline{Y}_{i\cdot\cdot}-\overline{Y}_{\cdot\cdot\cdot}\right)^2/(a-1)}{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk}-\overline{Y}_{i\cdot\cdot}\right)^2/\nu} \cdot \frac{a(bn-1)}{\nu} \tag{29}$$

In addition, adjusting for the distinct scaling of the distributions of the numerator (scaled by $n\sigma_\beta^2+\sigma^2$) and denominator (scaled by c) gives an approximate distribution of F_{anova}

$$F_{\text{anova}} \dot{\sim} \frac{a(bn-1)}{\nu}\frac{n\sigma_\beta^2+\sigma^2}{c}F(a-1,\nu,\lambda), \tag{30}$$

and inserting ν and c gives

$$F_{\text{anova}} \dot{\sim} \xi\, F(a-1,\nu,\lambda), \tag{31}$$

where

$$\xi = \frac{(bn-1)(n\sigma_\beta^2+\sigma^2)}{(b-1)(n\sigma_\beta^2+\sigma^2)+b(n-1)\sigma^2}. \tag{32}$$

According to (7) then the expected value of F_{anova} becomes

$$E(F_{\text{anova}}) \approx \xi\,\frac{\nu}{(a-1)(\nu-2)}\left(a-1+\frac{bn\sum_{i=1}^{a}\tau_i^2}{n\sigma_\beta^2+\sigma^2}\right) \tag{33}$$

and for sufficiently large ν

$$E(F_{\text{anova}}) \approx \xi \left(1 + \frac{bn\sum_{i=1}^{a}\tau_i^2}{(a-1)(n\sigma_\beta^2+\sigma^2)}\right) \tag{34}$$

which under H_0 reduces to

$$E(F_{\text{anova}}) \approx \xi. \tag{35}$$

When $\sigma_\beta^2 = 0$ then $\xi = 1$ and under H_0 then $E(F_{\text{anova}}) \approx 1$. For $\sigma_\beta^2 > 0$ then $\xi > 1$ implying that $E(F_{\text{anova}}) > 1$.

In practice, when data are analyzed with a one-way ANOVA the observed F_{anova}-statistic is (incorrectly) compared to a critical value obtained from an unscaled F-distribution, $F_{\alpha;a-1,a(bn-1)}$. The approximate type I error rate is found as

$$\text{Type I error rate} = 1 - G_s(F_{\alpha;a-1,a(bn-1)}; a-1, \nu), \tag{36}$$

where G_s refers to the scaled cumulative distribution function of F_{anova} given in (31) with $\lambda = 0$, since the type I error rate is defined under H_0.

Multiplying the scaled F-distribution by ξ^{-1} is a monotonic transformation (i.e. it preserves the order of the quantiles), hence the type I error rate can also be calculated as

$$\text{Type I error rate} \approx 1 - G\left(\xi^{-1} F_{\alpha;a-1,a(bn-1)}; a-1, \nu\right) \tag{37}$$

and the type I error rate can be found by means of a non-scaled F-distribution, which is readily available in most statistical software.

The type I error rate can also be expressed in terms of the variance components

$$\sigma^2_{\text{ratio}} = \frac{\sigma_\beta^2}{\sigma^2} \tag{38}$$

The effective degrees of freedom and the scaling factor is then found as

$$\begin{aligned}\nu &= \frac{a((b-1)(n\sigma_\beta^2+\sigma^2)+b(n-1)\sigma^2)^2}{(b-1)(n\sigma_\beta^2+\sigma^2)^2+b(n-1)(\sigma^2)^2} \cdot \frac{(\sigma^{-2})^2}{(\sigma^{-2})^2} \\ &= \frac{a((b-1)(n\sigma^2_{\text{ratio}}+1)+b(n-1))^2}{(b-1)(n\sigma^2_{\text{ratio}}+1)^2+b(n-1)}\end{aligned} \tag{39}$$

and

$$\begin{aligned}\xi &= \frac{(bn-1)(n\sigma_\beta^2+\sigma^2)}{(b-1)(n\sigma_\beta^2+\sigma^2)+b(n-1)\sigma^2} \cdot \frac{\sigma^{-2}}{\sigma^{-2}} \\ &= \frac{(bn-1)(n\sigma_{\text{ratio}}^2+1)}{(b-1)(n\sigma_{\text{ratio}}^2+1)+b(n-1)}. \end{aligned} \tag{40}$$

implying that the distribution of the F_{anova}-statistic and hence the type I error rate is influenced by the relative magnitudes of σ_β^2 and σ^2.

The type I error rate in special cases

In the special case where $n = 1$, that is, there is one slide per animal, then

$$\nu = a(b-1) \tag{41}$$

and

$$\gamma = 1 \tag{42}$$

For $\sigma_\beta^2 = 0$ then

$$\nu = a(bn-1) \tag{43}$$

and

$$\gamma = 1 \tag{44}$$

In both cases the approximate distribution in (31) becomes the ususal (appropriate) F-distribution and the type I error rate in (37) becomes α. This is what we expect since the hierarchical structure of the data in these cases will vanish so that model (19) becomes a suitable choice.

7.2.2 Results

Table 2 summarizes the type I error rate for different combinations of treatment groups, animals per treatment, slides per animal and ratios of the variance components. The levels of the first three factors, i.e. treatment groups, animals and slides were selected among actual levels identified in the literature study, although not every combination of the three factors occurred. From an earlier study (Hansen et al., 2014), where % tail DNA was used as an end point, $\hat{\sigma}_{\text{ratio}}^2 = 0.9 \approx 1$ and

Table 2: Type I error rate for different combinations of number of treatment groups (a), animals per treatment groups (b), slides per animal (n) and variance ratio $\sigma^2_{\text{ratio}} = \frac{\sigma^2_\beta}{\sigma^2}$. The effective denominator degrees of freedom ν, the scaling parameter ξ and the approximate $E(F)$ was calculated from (39), (40) and (33), respectively. The simulated type I error rate was based on 10000 simulations for each combination (each row) and the approximate type I error rate was found from (37). All approximate type I error rates were covered by the 95% confidence intervals for the simulated type I error rates except for two cases marked by asterisks. Type I error rates greater then 0.20 are marked in bold.

Treatment groups	Animals per treatment	Slides per animal	σ^2_{ratio}	Den DF $a(bn-1)$	Den DF ν	ξ	Approximate $E(F)$	Simulated type I error rate	Approximate type I error rate
2	4	2	0.5	14	12.50	1.40	1.67	0.094	0.094
2	4	2	1.0	14	10.90	1.61	1.98	0.118	0.120
2	4	2	2.0	14	9.14	1.84	2.36	0.150	0.148
2	4	3	0.5	22	17.96	1.77	2.00	0.139	0.137
2	4	3	1.0	22	14.29	2.20	2.56	0.186	0.183
2	4	3	2.0	22	10.85	2.65	3.25	**0.227**	**0.230**
2	8	2	0.5	30	26.89	1.36	1.47	0.090	0.092
2	8	2	1.0	30	23.69	1.55	1.70	0.120	0.114
2	8	2	2.0	30	20.21	1.74	1.94	0.140	0.138
2	8	3	0.5	46	37.56	1.72	1.81	0.131	0.133
2	8	3	1.0	46	30.25	2.09	2.24	0.178	0.174
2	8	3	2.0	46	23.54	2.48	2.71	**0.223**	**0.213***
6	4	2	0.5	42	37.50	1.40	1.48	0.147	0.149
6	4	2	1.0	42	32.71	1.61	1.72	**0.212**	**0.214**
6	4	2	2.0	42	27.42	1.84	1.99	**0.283**	**0.284**
6	4	3	0.5	66	53.89	1.77	1.84	**0.268**	**0.267**
6	4	3	1.0	66	42.86	2.20	2.31	**0.386**	**0.390**
6	4	3	2.0	66	32.55	2.65	2.83	**0.509**	**0.501**
6	8	2	0.5	90	80.67	1.36	1.40	0.149	0.144
6	8	2	1.0	90	71.07	1.55	1.60	0.194	**0.203***
6	8	2	2.0	90	60.62	1.74	1.80	**0.271**	**0.265**
6	8	3	0.5	138	112.69	1.72	1.75	**0.256**	**0.257**
6	8	3	1.0	138	90.75	2.09	2.14	**0.374**	**0.371**
6	8	3	2.0	138	70.61	2.48	2.55	**0.477**	**0.473**

the levels of the ratio were selected as 0.5, 1 and 2 times this approximate estimate.

The assumed denominator degrees of freedom was calculated as $a(bn-1)$ as this is used when data are fitted model (19). The effective degrees of freedom ν, the scaling factor ξ and the approximate $E(F)$ was calculated from (39), (40) and (33), respectively. The simulated type I error rate was obtained by simulating data structured as in Figure 1, and for each combination (each row in Table 2) 10000 simulations were conducted. The approximate type I error rates were calculated from (37). Throughout the nominal α was 0.05.

In all cases the assumed denominator degrees of freedom were greater than the effective denominator degrees of freedom, ν, and furthermore $\xi > 1$. This imposes additional skewness to the F-distribution implying a heavier right tail as seen in Figure 2, which illustrates the F-distributions for six treatment groups, four animals, three slides and $\sigma^2_{\text{ratio}} = 1$.

Increasing the number of treatment groups enhanced the type I error rate considerably. The same was in evidence when the number of slides per animal were increased. Interestingly, the number of animals per treatment group did not affect the type I error rate noticeably. Increasing σ^2_{ratio} (increasing σ^2_β relative to σ^2) in general increased the type I error rate. All cases resulted in a type I error rate greater than the nominal α at 0.05. Most combinations gave type I error rates greater than 0.10 and almost half resulted in type I error rates greater than 0.20.

The validity of the approximate type I error rates was assessed by making an informal comparison to the simulated type I error rates. To quantify the simulation uncertainty the standard errors were calculated as $\text{se}(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}$ and Wald based 95% confidence intervals (CI) were obtained (not shown). The simulated type I error rates were between 0.090 and 0.509, hence the standard errors were between 0.003 and 0.005. In all but two cases the approximate type I error rates were covered by the CI for the simulated type I error rates. This agrees with the expectation of 1 to 2 values falling outside the CI given the number of comparisons and the confidence level. The two cases not covered by the CI are marked with asterisks in Table 2. A 99% CI for the simulated type I error rates covered all the approximate type I error rates.

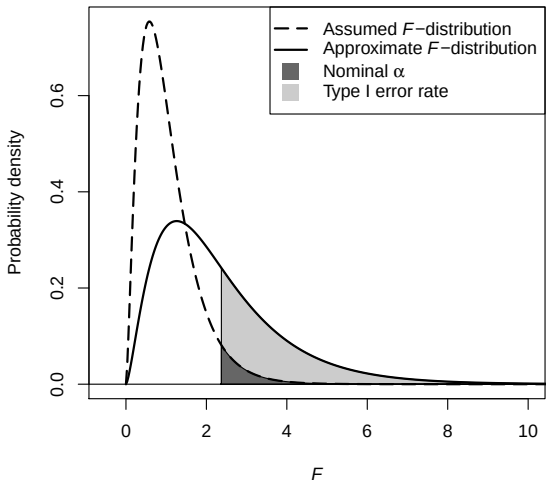

Figure 2: The F-distributions in case of six treatment groups, four animals per treatment, three slides per animal and $\sigma^2_{\text{ratio}} = 1$. The assumed F-distribution refers to the distribution from which the critical value is obtained. The approximate F-distribution is the distribution of F_{anova} as defined in (31). The approximate F-distribution has a heavier right tail implying that the type I error rate is greater than the nominal α at 0.05.

8 Discussion

This study aimed at addressing potential issues concerning the analysis of Comet assay data. First, from the literature study it was not possible to deduce exactly how data were analyzed, which impedes reproducibility and blurs the interpretation of the reported results. Even if some researchers analyze data properly, we find it likely that others (e.g. new researchers in the field) may be inspired by the insufficient description of the statistical modeling in the papers and thereby may fail to allow properly for the nested structure. Second, as we suspect that the nested structure in data is not accounted for in the statistical model we investigated the implications in terms of the type I error rate. Approximate formulas for one likely case were derived to examine in which way the type I error rate was affected.

Type I error rates for different combinations of the factors as they appeared in the literature study demonstrated that the inflation is in fact non-trivial. When the

cell scores are used as the response all type I error rates examined in the current study were severely inflated yielding type I error rates as high as 0.950. These results were seen for combination of factors as they appeared in the literature study and we therefore consider these results likely to occur in practice. Interestingly, the variance components reflecting the animal and slide variation were relatively small compared to the residual variation, i.e. the ratios were $\frac{\sigma^2_\beta}{\sigma^2} = 0.026$ and $\frac{\sigma^2_\gamma}{\sigma^2} = 0.013$, respectively. Even so, the results show that the hypothesis test yields completely unreliable results from which erroneous inferences are made. This means that even factors that contribute with variation that seem negligible can have a huge impact on the results. One reason may be the high number of scored cells, which often in practice is 50 or 100 cells per slide.

Closed-form expressions were derived for the case where a summary statistic is calculated for each slide and they showed that the actual sampling distribution approximately follows a scaled F-distribution. Both the number of treatment groups, animals per treatment, slides per animal, the variance ratio $\sigma^2_{\text{ratio}} = \frac{\sigma^2_\beta}{\sigma^2}$ and the significance level, α, influences the shape of this distribution and hence the type I error rate. For the cases shown here the approximate type I error rates were between 0.094 and 0.501, and for all combinations they were greater than the nominal α at 0.05. Almost half of the cases resulted in type I error rates greater than 0.20. In practice, the number of animals did not seem to have a noticeable effect on the type I error rate but all other factors that appeared in the closed-form expressions affected the type I error rate appreciably.

Our objective was to illustrate the implications in a simple manner with the hope of motivating researchers within the field to reconsider the statistical modeling. As the design considered here is widespread across various scientific areas we believe that the results may be equally relevant to researchers in other fields.

Appendices

A Expectation and variance of Y_{ijk}, $\overline{Y}_{ij\cdot}$, $\overline{Y}_{i\cdot\cdot}$ and $\overline{Y}_{\cdot\cdot\cdot}$

In the following the expectation and variance of Y_{ijk}, $\overline{Y}_{ij\cdot}$, $\overline{Y}_{i\cdot\cdot}$ and $\overline{Y}_{\cdot\cdot\cdot}$ is derived from model (2). All terms in the model are assumed to be independent and the following results are used:

Let $X_1, X_2, ..., X_n$ be random variables and let $T = \sum_{i=1}^{n} a_i X_i$. Then

$$E(T) = \sum_{i=1}^{n} a_i E(X_i) \tag{45}$$

and if $X_1, X_2, ..., X_n$ are independent then

$$\mathrm{Var}(T) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i) \tag{46}$$

(Hogg et al., 2005)

Expectation and variance of Y_{ijk}

Given that

$$Y_{ijk} = \mu + \tau_i + \beta_{j(i)} + \varepsilon_{(ij)k} \tag{47}$$

then

$$\begin{aligned} E(Y_{ijk}) &= E(\mu) + E(\tau_i) + E\left(\beta_{j(i)}\right) + E\left(\varepsilon_{(ij)k}\right) \\ &= \mu + \tau_i \end{aligned} \tag{48}$$

and

$$\begin{aligned} \mathrm{Var}(Y_{ijk}) &= \mathrm{Var}(\mu) + \mathrm{Var}(\tau_i) + \mathrm{Var}\left(\beta_{j(i)}\right) + \mathrm{Var}\left(\varepsilon_{(ij)k}\right) \\ &= \sigma_\beta^2 + \sigma^2 \end{aligned} \tag{49}$$

Expectation and variance of $\overline{Y}_{ij\cdot}$

The group mean $\overline{Y}_{ij\cdot}$ is obtained as

$$\begin{aligned}\overline{Y}_{ij\cdot} &= \frac{1}{n}\sum_{k=1}^{n} Y_{ijk} \\ &= \frac{1}{n}\sum_{k=1}^{n} \left(\mu + \tau_i + \beta_{j(i)} + \varepsilon_{(ij)k}\right) \\ &= \mu + \tau_i + \beta_{j(i)} + \frac{1}{n}\sum_{k=1}^{n} \varepsilon_{(ij)k}. \end{aligned} \tag{50}$$

Then

$$\begin{aligned} E(\overline{Y}_{ij\cdot}) &= E(\mu) + E(\tau_i) + E\left(\beta_{j(i)}\right) + E\left(\frac{1}{n}\sum_{k=1}^{n} \varepsilon_{(ij)k}\right) \\ &= \mu + \tau_i \end{aligned} \tag{51}$$

and

$$\begin{aligned} \mathsf{Var}(\overline{Y}_{ij\cdot}) &= \mathsf{Var}(\mu) + \mathsf{Var}(\tau_i) + \mathsf{Var}\left(\beta_{j(i)}\right) + \mathsf{Var}\left(\frac{1}{n}\sum_{k=1}^{n} \varepsilon_{(ij)k}\right) \\ &= \sigma_\beta^2 + \frac{\sigma^2}{n} \\ &= \frac{n\sigma_\beta^2 + \sigma^2}{n} \end{aligned} \tag{52}$$

Expectation and variance of $\overline{Y}_{i\cdot\cdot}$

The group mean $\overline{Y}_{i\cdot\cdot}$ is obtained as

$$\begin{aligned}\overline{Y}_{i\cdot\cdot} &= \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk} \\ &= \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\mu + \tau_i + \beta_{j(i)} + \varepsilon_{(ij)k}\right) \\ &= \mu + \tau_i + \frac{1}{b}\sum_{j=1}^{b}\beta_{j(i)} + \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{(ij)k}. \end{aligned} \tag{53}$$

Then

$$\begin{aligned}E(\overline{Y}_{i\cdot\cdot}) &= E(\mu) + E(\tau_i) + E\left(\frac{1}{b}\sum_{j=1}^{b}\beta_{j(i)}\right) + E\left(\frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{(ij)k}\right) \\ &= \mu + \tau_i \end{aligned} \tag{54}$$

and

$$\begin{aligned}\mathsf{Var}(\overline{Y}_{i\cdot\cdot}) &= \mathsf{Var}(\mu) + \mathsf{Var}(\tau_i) + \mathsf{Var}\left(\frac{1}{b}\sum_{j=1}^{b}\beta_{j(i)}\right) + \mathsf{Var}\left(\frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{(ij)k}\right) \\ &= \frac{\sigma_\beta^2}{b} + \frac{\sigma^2}{bn} \\ &= \frac{n\sigma_\beta^2 + \sigma^2}{bn} \end{aligned} \tag{55}$$

Expectation and variance of $\overline{Y}_{...}$

The group mean $\overline{Y}_{...}$ is obtained as

$$\begin{aligned}\overline{Y}_{...} &= \frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk} \\ &= \frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\mu+\tau_i+\beta_{j(i)}+\varepsilon_{(ij)k}\right) \\ &= \mu+\frac{1}{a}\sum_{i=1}^{a}\tau_i+\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\beta_{j(i)}+\frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{(ij)k}. \end{aligned} \tag{56}$$

Then

$$\begin{aligned}E(\overline{Y}_{...}) &= E(\mu)+E\left(\frac{1}{a}\sum_{i=1}^{a}\tau_i\right)+E\left(\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\beta_{j(i)}\right) \\ &\quad + E\left(\frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{(ij)k}\right) \\ &= \mu \end{aligned} \tag{57}$$

and

$$\begin{aligned}\mathsf{Var}(\overline{Y}_{...}) &= \mathsf{Var}(\mu)+\mathsf{Var}\left(\frac{1}{a}\sum_{i=1}^{a}\tau_i\right)+\mathsf{Var}\left(\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\beta_{j(i)}\right) \\ &\quad + \mathsf{Var}\left(\frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{(ij)k}\right) \\ &= \frac{\sigma_\beta^2}{ab}+\frac{\sigma^2}{abn} \\ &= = \frac{n\sigma_\beta^2+\sigma^2}{abn} \end{aligned} \tag{58}$$

B Distribution of the sum of squares

In the following the distributions of the relevant sum of squares that appear in the F_{anova}-statistic presented in section 6 are derived. The results are based on the

definition of model (2) and the results obtained in appendix A.

Distribution of $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk}-\overline{Y}_{ij\cdot})^2$

According to (2) and (50) then

$$\begin{aligned} Y_{ijk}-\overline{Y}_{ij\cdot} &= \mu+\tau_i+\beta_{(i)j}+\varepsilon_{(ij)k}-\left(\mu+\tau_i+\beta_{(i)j}+\frac{1}{n}\sum_{k=1}^{n}\varepsilon_{(ij)k}\right) \\ &= \varepsilon_{(ij)k}-\frac{1}{n}\sum_{k=1}^{n}\varepsilon_{(ij)k} \end{aligned} \tag{59}$$

Since $\varepsilon_{(ij)k}\sim N(0,\sigma^2)$ and $\frac{1}{n}\sum_{k=1}^{n}\varepsilon_{(ij)k}\sim N\left(0,\frac{\sigma^2}{n}\right)$ then

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\varepsilon_{(ij)k}-\frac{1}{n}\sum_{k=1}^{n}\varepsilon_{(ij)k}\right)^2\sim\sigma^2\chi^2\left(ab(n-1)\right) \tag{60}$$

hence

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk}-\overline{Y}_{ij\cdot}\right)^2\sim\sigma^2\chi^2\left(ab(n-1)\right) \tag{61}$$

Distribution of $n\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{Y}_{ij\cdot}-\overline{Y}_{i\cdot\cdot})^2$

According to (50) and (53) then

$$\begin{aligned} Y_{ij\cdot}-\overline{Y}_{i\cdot\cdot} &= \mu+\tau_i+\beta_{(i)j}+\frac{1}{n}\sum_{k=1}^{n}\varepsilon_{(ij)k} \\ &\quad-\left(\mu+\tau_i+\frac{1}{b}\sum_{j=1}^{b}\beta_{(i)j}+\frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{(ij)k}\right) \\ &= \beta_{(i)j}+\frac{1}{n}\sum_{k=1}^{n}\varepsilon_{(ij)k}-\frac{1}{b}\sum_{j=1}^{b}\left(\beta_{(i)j}+\frac{1}{n}\sum_{k=1}^{n}\varepsilon_{(ij)k}\right) \end{aligned} \tag{62}$$

where the last term is seen to an average of the first two terms. Also,

$$\beta_{(i)j} + \frac{1}{n}\sum_{k=1}^{n}\varepsilon_{(ij)k} \sim N\left(0, \frac{n\sigma_\beta^2 + \sigma^2}{n}\right), \tag{63}$$

and

$$\frac{1}{b}\sum_{j=1}^{b}\left(\beta_{(i)j} + \frac{1}{n}\sum_{k=1}^{n}\varepsilon_{(ij)k}\right) \sim N\left(0, \frac{n\sigma_\beta^2 + \sigma^2}{bn}\right) \tag{64}$$

so that

$$n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot}\right)^2 \sim (n\sigma_\beta^2 + \sigma^2)\chi^2\left(a(b-1)\right) \tag{65}$$

Distribution of $bn\sum_{i=1}^{a}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots})^2$

According to (53) and (56) then

$$\begin{aligned}
Y_{i\cdot\cdot} - \overline{Y}_{\cdots} &= \mu + \tau_i + \frac{1}{b}\sum_{j=1}^{b}\beta_{(i)j} + \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{(ij)k} \\
&\quad - \left(\mu + \frac{1}{a}\sum_{i=1}^{a}\tau_i + \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\beta_{(i)j} + \frac{1}{abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{(ij)k}\right) \\
&= \tau_i + \frac{1}{b}\sum_{j=1}^{b}\beta_{(i)j} + \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{(ij)k} \\
&\quad - \frac{1}{a}\sum_{i=1}^{a}\left(\tau_i + \frac{1}{b}\sum_{j=1}^{b}\beta_{(i)j} + \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{(ij)k}\right)
\end{aligned} \tag{66}$$

where the last term is seen to be an average of the first three terms. Also,

$$\tau_i + \frac{1}{b}\sum_{j=1}^{b}\beta_{(i)j} + \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{(ij)k} \sim N\left(\tau_i, \frac{n\sigma_\beta^2 + \sigma^2}{bn}\right), \tag{67}$$

and

$$\frac{1}{a}\sum_{i=1}^{a}\left(\tau_i + \frac{1}{b}\sum_{j=1}^{b}\beta_{(i)j} + \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{(ij)k}\right) \sim N\left(0, \frac{n\sigma_\beta^2 + \sigma^2}{abn}\right), \tag{68}$$

so that

$$bn\sum_{i=1}^{a}\left(\overline{Y}_{i\cdot\cdot}-\overline{Y}_{\cdot\cdot\cdot}\right)^2\sim(n\sigma_\beta^2+\sigma^2)\chi^2\left(a-1,\lambda\right), \tag{69}$$

where

$$\lambda=\frac{bn\sum_{i=1}^{n}\tau_i^2}{n\sigma_\beta^2+\sigma^2} \tag{70}$$

C Approximate distribution of a linear combination of χ^2 variates

The sum of squares in the denominator of (21) can be partitioned as

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk}-\overline{Y}_{i\cdot\cdot}\right)^2=n\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\overline{Y}_{ij\cdot}-\overline{Y}_{i\cdot\cdot}\right)^2+\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk}-\overline{Y}_{ij\cdot}\right)^2 \tag{71}$$

which can be expressed as

$$V=V_1+V_2, \tag{72}$$

where $V_i\sim c_i\chi^2(\nu_i), i=1,...,2$. An exact distribution of V is given in Box (1954) and Satterthwaite (1941). However, a more accessible representation can be accomplished by means of the Welch-Satterthwaite approach, where the distribution of V is approximated by a scaled χ^2-distribution. The scaling factor and the degrees of freedom of the χ^2-distribution is found by matching the first two moments of V and the approximate distribution. In the following it will be utilized that $E(\chi^2_m)=m$ and Var$(\chi^2_m)=2m$ (Dobson, 2002; Johnson et al., 1994).

First, the distribution of V is approximated with a scaled χ^2-distribution of the form

$$V\dot{\sim}c\chi^2(\nu). \tag{73}$$

By equating the first two moments of V and $c\chi^2(\nu)$ we get $E(V_1+V_2)=E(c\chi^2(\nu))$, so that

$$c_1\nu_1+c_2\nu_2=c\nu. \tag{74}$$

Since V_1 and V_2 are independent (which can be shown using Fisher-Cochran's Theorem (Rao, 1973)), then $\text{Var}(V_1 + V_2) = \text{Var}(c\chi^2(\nu))$, so that

$$2c_1^2\nu_1 + 2c_2^2\nu_2 = 2c^2\nu. \tag{75}$$

The scaling factor, c, is found by inserting (74) into (75)

$$2c_1^2\nu_1 + 2c_2^2\nu_2 = 2c(c_1\nu_1 + c_2\nu_2) \tag{76}$$

so that

$$c = \frac{c_1^2\nu_1 + c_2^2\nu_2}{c_1\nu_1 + c_2\nu_2}. \tag{77}$$

The degrees of freedom, ν, is obtained by inserting (77) in (74) and rearranging, so that

$$\nu = \frac{(c_1\nu_1 + c_2\nu_2)^2}{c_1^2\nu_1 + c_2^2\nu_2}. \tag{78}$$

References

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics 25*(2), 290–302.

Bright, J., M. Aylott, S. Bate, H. Geys, P. Jarvis, J. Saul, and R. Vonk (2011). Recommendations on the statistical analysis of the comet assay. *Pharmaceutical Statistics 10*, 485–493.

Dobson, A. J. (2002). *An introduction to generalized linear models* (second ed.). Texts in Statistical Science. USA: Chapman and Hall/CRC.

Duez, P., G. Dehon, A. Kumps, and J. Dubois (2003). Statistics of the comet assay: a key to discriminate between genotoxic effects. *Mutagenesis 18*(2), 159–166.

Ejchart, A. and N. Sadlej-Sosnowska (2003). Statistical evaluation and comparison of comet assay results. *Mutation Research 534*, 85–92.

Hansen, M. K., A. K. Sharma, M. Dybdahl, J. Boberg, and M. Kulahci (2014). *In vivo* comet assay - statistical analysis and power calculations of mice testicular cells. Submitted.

Hartmann, A., E. Agurell, C. Beevers, S. brendler Schwaab, B. Burlinson, P. Clay, A. Collins, A. Smith, G. Speit, V. Thybaud, and R. R. Tice (2003). Recommendations for conducting the *in vivo* alkaline Comet assay. *Mutagenesis 18*(1), 45–51.

Hogg, R. V., J. W. McKean, and A. T. Craig (2005). *Introduction to Mathematical Statistics* (sixth ed.). USA: Prentice Hall.

Jeggo, P. A. and M. Löbrich (2007). Dna double-strand breaks: their cellular and clinical impact? *Oncogene 26*, 7717–7719.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1994). *Continuous univariate distributions* (second ed.), Volume 1 of *Wiley series in probability and mathematical statistics*. John Wiley & Sons.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1995). *Continuous univariate distributions* (second ed.), Volume 2 of *Wiley series in probability and mathematical statistics*. John Wiley & Sons.

Johnson, R., J. Freund, and I. Miller (2010). *Probability and Statistics for Engineers* (eight ed.). Prentice Hall.

Kumaravel, T. S. and A. N. Jha (2006). Reliable Comet assay measurements for detecting dna damage induced by ionising radiation and chemicals. *Mutation Research 605*, 7–16.

Lovell, D. P. and T. Omori (2008). Statistical issues in the use of the comet assay. *Mutagenesis 23*(3), 171–182.

Lovell, D. P., G. Thomas, and R. Dubow (1999). Issues related to the experimental design and subsequent statistical analysis of in vivo and in vitro comet studies. *Teratogenesis, Carcinogenesis, and Mutagenesis 19*, 109–119.

Malta, L. G., F. G. Ghiraldini, R. Reis, M. do Vale Oliveira, L. B. Silva, and G. M. Pastore (2012). In vivo analysis of antigenotoxic and antimutagenic properties of two brazilian cerrado fruits and the identification of phenolic phytochemicals. *Food Research International 49*, 604–611.

Montgomery, D. C. (2005). *Design and Analysis of Experiments* (sixth ed.). USA: John Wiley & Sons, Inc.

Olive, P. L., J. P. Banáth, and R. E. Durand (1990). Heterogeneity in radiation-induced dna damage and repair in tumor and normal cells measured using the "Comet" assay. *Radiation Research 122*(1), 86–94.

Pesarini, J. R., P. T. Zaninetti, M. O. Mauro, C. M. Carreira, J. B. Dichi, L. R. Ribeiro, M. S. Mantovani, and R. J. Oliveira (2013). Antimutagenic and anticarcinogenic effects of wheat bran *in vivo*. *Genetics and Molecular Research 12*(2), 1646–1659.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications* (second ed.). Wiley series in probability and mathematical statistics. USA: John Wiley & Sons.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika 6*(5), 309–316.

Verde, P. E., L. A. Geracitano, L. L. Amado, C. E. Rosa, A. Bianchini, and J. M. Monserrat (2006). Application of public-domain statistical analysis software for evaluation and comparison of comet assay data. *Mutation Research 604*, 71–82.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika 29*(3/4), 350–362.

Wiklund, S. J. and E. Agurell (2003). Aspects of design and statistical analysis in the comet assay. *Mutagenesis 18*(2), 167–175.

Zan, M. A., A. B. F. Ferraz, M. F. Richter, J. N. Picada, H. H. R. de Andrade, M. Lehmann, R. R. Dihl, E. Nunes, J. Semedo, and J. D. Silva (2013). *In vivo* genotoxicity evaluation of an artichoke (*Cynara scolymus* l.) aqueous extract. *Journal of Food Science 78*(2), T367–T371.

Appendix C

On the Type I Error Rate When the Hierarchical Structure of Data Is Ignored

---

Hansen, M. K. and Kulahci, M. (2014). On the Type I Error Rate When the Hierarchical Structure of Data Is Ignored. *The American Statistician*, submitted.

On the Type I Error Rate When the Hierarchical Structure of Data Is Ignored

Abstract

Hierarchical data arise naturally in many different fields. The statistical analysis should suitably accommodate the hierarchical structure. This sometimes however seems to be disregarded, which leads to considerable impact on the type I error rate. For different combinations of relevant factors the type I error rates are between 0.074 and ≈ 1, i.e. for all the combinations the type I error rate is greater than the nominal α at 0.05, and in most cases the inflation is substantial. Closed-form expressions for the approximate type I error rate are provided to clarify how they are affected. With this study we aim to show the inferential implications that result from disregarding the hierarchical structure in the hope of motivating researchers not to underestimate the severe consequences and to properly account for the hierarchical structure.

KEY WORDS: Inflated alpha, nested design, violation of assumptions, non-independence, linear mixed-effects model

1 INTRODUCTION

Data with hierarchical structure are encountered in many different fields of research and application. Among others, these studies appear frequently in phychology, sociology, toxicology, pharmacology, neurology etc. In particular, hierarchical data are often the outcome from experimental studies involving animals and humans. This work was inspired by our practical research on the design and analysis of animal experiments.

Data with a hierarchical structure are naturally accommodated by hierarchical models, also known as multilevel models, nested models, mixed-effects models or random-effects models (Laird and Ware 1982; Raudenbush and Bryk 2002; Gelman and Hill 2007; Verbeke and Molenberghs 2009). However, in some fields of research these types of models seem not to have gained any footing. For illustration purposes we will focus on a simple and in our experience quite common hierarchical structure shown in Figure 1. Here a number of treatments are of interest and within each treatment the observations are naturally clustered in groups. In this case various statistical approaches are feasible depending on the aim of the study and one popular choice is to apply a one-way ANOVA. For data arising from the design in Figure 1, a one-way ANOVA is certainly a possibility if data are suitably aggregated and if the inference of interest is at the aggregate level. However, this approach precludes the possibility of making inference at the observational level as this would lead to what is known as an ecological fallacy (Robinson 1950; Freedman

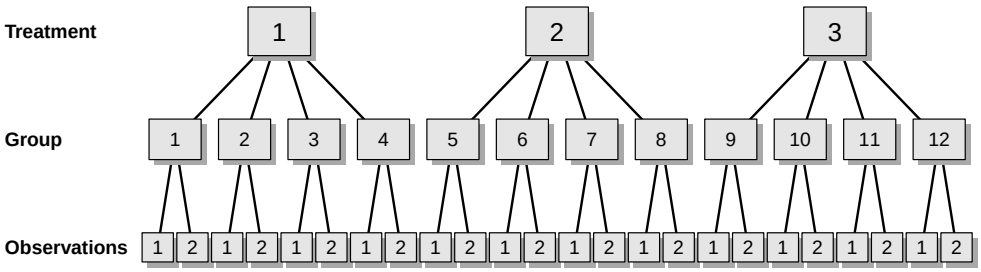

Figure 1: Outline of a common experimental design. This example shows three treatments, four groups per treatment and two observations per group.

2001; Gelman et al. 2008)

A main concern is that the statistical model should accommodate the data to which it is applied. Thus, even though a one-way ANOVA sometimes can be a valid choice for aggregated hierarchical data, it is in general not suitable for raw hierarchical data. The most crucial assumption of a one-way ANOVA is that the observations are independent, which is obviously not met when data are hierarchically structured. Unfortunately, the use of a one-way ANOVA in the analysis of hierarchical data sometimes occurs. This issue has been addressed by various authors (Kenny and Judd 1986; Holson and Pearce 1992; Kromrey and Dickinson 1996; Tasca et al. 2009; McCoach and Adelson 2010), yet the occasional lack of consideration toward the dependency among observations seems to persist (Wampold and Serlin 2000; Baldwin et al. 2005; Baccaglini et al. 2010; Musca et al. 2011; Hansen and Kulahci 2014).

The aim of this paper is to yet again raise awareness about this issue and shed light on the implications when the hierarchical structure is ignored for one commonly occuring type of hierarchical data. The implications are

quantified in terms of the type I error rate and closed-form expressions for the approximate type I error rate for balanced data are derived to examine the source of the impact on this measure. Type I error rates for different ratios of the between-group and within-group variation and for various sample sizes are calculated from the closed-form expressions, which are validated by simulations. It is our hope that this exposition can add to the appreciation of properly accounting for the non-independence among observations stemming from the hierarchical structure of the data.

2 EXPERIMENTAL DESIGN AND STATISTICAL MODELS

One type of hierarchical structure that is commonly seen is illustrated in Figure 1. There are numerous examples where this structure arises naturally, e.g. when teaching methods are compared on students that are clustered within classes, when various doses of a given compound is tested in multiple offsprings per litter or when observations repeatedly are collected from each subject. This structure is thus a natural outcome in many settings, and in the following, we use it to illustrate the pitfalls of ignoring dependency among observations.

One way to analyze data structured as in Figure 1 is to fit a linear mixed-effects model with treatment as a fixed effect and group as a random effect and with group nested within treatment. Hence a proper model is

$$Y_{ijk} = \mu + \tau_i + \beta_{(i)j} + \varepsilon_{(ij)k} \tag{1}$$

where

$$i = 1, ..., a,\ j = 1, ..., b,\ k = 1, ..., n,$$

$$\beta_{(i)j} \sim N(0, \sigma^2_\beta), \quad \varepsilon_{(ij)k} \sim N(0, \sigma^2).$$

Y_{ijk} is the ijkth observation and μ and τ_i are the fixed effects for the intercept and treatment, respectively. $\beta_{(i)j}$ is the random effect of the jth group nested within the ith treatment and $\varepsilon_{(ij)k}$ is the within-group error. The parentheses in the subscripts indicate the nesting structure with the parent level(s) given inside the parentheses (Montgomery 2005).

In some cases, hierarchical data are reported to be analyzed with a one-way ANOVA. The model underlying this method is typically:

$$Y_{ij^*} = \mu + \tau_i + \varepsilon_{ij^*} \tag{2}$$

where $i = 1, ..., a,\ j^* = 1, ..., bn$ and $\varepsilon_{ij^*} \sim N(0, \sigma^{*2})$. Using model (2) invokes the assumption that all observations are independent. However, for hierarchical data structured as in Figure 1 the observations are in fact not independent. Observations within each group are correlated and if not accounted for in the statistical modeling, it inflicts severe implications with regard to the inference that is made about the treatment parameter.

3 NOTATION

Let $Y_{ij.} = \sum_{k=1}^{n} Y_{ijk}$, $Y_{i..} = \sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk}$, $Y_{...} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} Y_{ijk}$ and let $\overline{Y}_{ij.} = \frac{1}{n}Y_{ij.}$, $\overline{Y}_{i..} = \frac{1}{bn}Y_{i..}$, $\overline{Y}_{...} = \frac{1}{abn}Y_{...}$. If $V \sim c\chi^2(\nu, \lambda)$ then V follows a scaled non-central χ^2-distribution with ν degrees of freedom, scaling parameter c and non-centrality parameter λ. If $c = 1$ and $\lambda = 0$ then we say that V follows a non-scaled central χ^2-distribution. If $W \sim cF(\nu_1, \nu_2, \lambda)$ then W has a scaled non-central F-distribution with ν_1 and ν_2 degrees of freedom, scaling parameter c and non-centrality parameter λ. The cumulative distribution function of W evaluated at w is denoted $G(w; \nu_1, \nu_2, \lambda)$ when W follows a non-scaled distribution or $G_s(w; \nu_1, \nu_2, \lambda)$ if the distribution is scaled. If $W \sim F(\nu_1, \nu_2)$ then W has a non-scaled central F-distribution with the critical value $F_{\alpha;\nu_1,\nu_2}$ being the $(1-\alpha)$th quantile such that $G(F_{\alpha;\nu_1,\nu_2}; \nu_1, \nu_2) = 1-\alpha$.

4 THE TYPE I ERROR RATE

In some studies (e.g. Morales et al. 2013; Chekhun et al. 2013; Almeida et al. 2013; Kelly et al. 2013) it seems that, rather than fitting a model that takes into account the hierarchical structure as in (1), a fixed-effects model such as given in (2) is employed

$$Y_{ij^*} = \mu + \tau_i + \varepsilon_{ij^*}$$

where $i = 1, ..., a$, $j^* = 1, ..., bn$ and $\varepsilon_{ij^*} \sim N(0, \sigma^{*2})$. This model typi-

cally underlies what is referred to as a one-way ANOVA. The F-statistic is calculated as

$$F_{\text{anova}} = \frac{bn \sum_{i=1}^{a} \left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2 / (a-1)}{\sum_{i=1}^{a} \sum_{j^*=1}^{bn} \left(Y_{ij^*} - \overline{Y}_{i\cdot}\right)^2 / (a(bn-1))}$$

which is expressed within the framework of model (1) as

$$F_{\text{anova}} = \frac{bn \sum_{i=1}^{a} \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot}\right)^2 / (a-1)}{\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} \left(Y_{ijk} - \overline{Y}_{i\cdot\cdot}\right)^2 / (a(bn-1))} \tag{3}$$

The denominator of (3) can be rewritten as

$$\left\{ n \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot}\right)^2 + \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} \left(Y_{ijk} - \overline{Y}_{ij\cdot}\right)^2 \right\} / \left(a(b-1) + ab(n-1)\right) \tag{4}$$

implying that sum of squares and the degrees of freedom in the denominator is attributable both to the grouping and the error part.

A nice feature of fitting data with model (1) is that the sums of squares in the numerator and denominator of the relevant F-statistic both follow χ^2-distributions that are scaled by $n\sigma_\beta^2 + \sigma^2$, that is, they cancel out and the ratio follows a central non-scaled F-distribution. This is not the case for the F_{anova}-statistic in (3) as the sums of squares follow χ^2-distributions that are

scaled differently. The sum of squares in the numerator is distributed as

$$bn \sum_{i=1}^{a} \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot} \right)^2 \sim \left(n\sigma_\beta^2 + \sigma^2 \right) \chi^2 \left(a - 1, \lambda \right),$$

where

$$\lambda = \frac{bn \sum_{i=1}^{a} \tau_i^2}{n\sigma_\beta^2 + \sigma^2}.$$

Since Y_{ijk} are not independent the sum of squares in the denominator of (3) does not follow the usual χ^2-distribution (see Box (1954) for details). However, looking separately at the two terms in the numerator of (4) gives

$$n \sum_{i=1}^{a} \sum_{j=1}^{b} \left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} \right)^2 \sim \left(n\sigma_\beta^2 + \sigma^2 \right) \chi^2(a(b-1)),$$

and

$$\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} \left(Y_{ijk} - \overline{Y}_{ij\cdot} \right)^2 \sim \sigma^2 \chi^2(ab(n-1)),$$

that is, $\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} \left(Y_{ijk} - \overline{Y}_{i\cdot\cdot} \right)^2$ is a linear combination of independent χ^2-distributed random variables. An approximate distribution is obtained using the rationale of the Welch-Satterthwaite approximation (Welch 1938; Satterthwaite 1941; Box 1954). The sum of squares is approximated by a scaled χ^2-distribution

$$\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} \left(Y_{ijk} - \overline{Y}_{i\cdot\cdot} \right)^2 \doteqdot c\chi^2(\nu)$$

and c and ν are found by matching the first two moments, so that

$$c = \frac{(b-1)(n\sigma_\beta^2+\sigma^2)^2 + b(n-1)(\sigma^2)^2}{(b-1)(n\sigma_\beta^2+\sigma^2) + b(n-1)\sigma^2}$$

and

$$\nu = \frac{a((b-1)(n\sigma_\beta^2+\sigma^2) + b(n-1)\sigma^2)^2}{(b-1)(n\sigma_\beta^2+\sigma^2)^2 + b(n-1)(\sigma^2)^2},$$

where ν is known as the effective degrees of freedom (Satterthwaite 1941). A ratio of χ^2-distributed random variables each divided by its degrees of freedom is F-distributed. However, the sum of squares in the denominator of (3) is not divided by its effective degrees of freedom ν but by $a(bn-1)$, so that

$$F_{\text{anova}} = \frac{bn\sum_{i=1}^{a}\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot}\right)^2/(a-1)}{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk} - \overline{Y}_{i\cdot\cdot}\right)^2/\nu} \cdot \frac{a(bn-1)}{\nu}$$

In addition, adjusting for the different scaling of the distributions of the numerator (scaled by $n\sigma_\beta^2+\sigma^2$) and the denominator (scaled by c) gives an approximate distribution of F_{anova}

$$F_{\text{anova}} \dot{\sim} \frac{a(bn-1)}{\nu}\frac{n\sigma_\beta^2+\sigma^2}{c}F(a-1,\nu,\lambda),$$

and inserting ν and c gives

$$F_{\text{anova}} \dot{\sim} \xi\, F(a-1,\nu,\lambda), \tag{5}$$

where

$$\xi = \frac{(bn-1)(n\sigma_\beta^2+\sigma^2)}{(b-1)(n\sigma_\beta^2+\sigma^2)+b(n-1)\sigma^2}.$$

Since the expected value of an F-distributed random variable with ν_1 and ν_2 degrees of freedom is $E(F) = \frac{\nu_2(\nu_1+\lambda)}{\nu_1(\nu_2-2)}$ (Johnson et al. 1995), then

$$E(F_{\text{anova}}) \approx \xi \frac{\nu}{(a-1)(\nu-2)}\left(a-1+\frac{bn\sum_{i=1}^{a}\tau_i^2}{n\sigma_\beta^2+\sigma^2}\right)$$

and for sufficiently large ν

$$E(F_{\text{anova}}) \approx \xi \left(1+\frac{bn\sum_{i=1}^{a}\tau_i^2}{(a-1)(n\sigma_\beta^2+\sigma^2)}\right)$$

which under the null hypothesis (H_0) that $\tau_1 = \cdots = \tau_a = 0$ reduces to

$$E(F_{\text{anova}}) \approx \xi.$$

When $\sigma_\beta^2 = 0$ then $\xi = 1$ and under H_0 then $E(F_{\text{anova}}) \approx 1$. For $\sigma_\beta^2 > 0$ then $\xi > 1$ implying that $E(F_{\text{anova}}) > 1$.

In practice, when data are analyzed with one-way ANOVA, the observed F_{anova}-statistic is (incorrectly) compared to a critical value obtained from an unscaled F-distribution, $F_{\alpha;a-1,a(bn-1)}$. The approximate type I error rate is

found as

$$\text{Type I error rate} \approx 1 - G_s(F_{\alpha;a-1,a(bn-1)}; a-1, \nu),$$

where G_s refers to the scaled cumulative distribution function of F_{anova} given in (5) with $\lambda = 0$, since the type I error rate is defined under H_0.

Multiplying the scaled F-distribution by ξ^{-1} is a monotonic transformation, hence the type I error rate can also be calculated as

$$\text{Type I error rate} \approx 1 - G\left(\xi^{-1}\, F_{\alpha;a-1,a(bn-1)}; a-1, \nu\right) \tag{6}$$

and the type I error rate can be obtained by means of a non-scaled F-distribution, which is readily available in most statistical software.

The type I error rate can also be expressed in terms of the ratio of the variance components

$$\sigma^2_{\text{ratio}} = \frac{\sigma^2_\beta}{\sigma^2}$$

The effective degrees of freedom and the inverse scaling factor is then found as

$$\nu = \frac{a((b-1)(n\sigma^2_{\text{ratio}}+1) + b(n-1))^2}{(b-1)(n\sigma^2_{\text{ratio}}+1)^2 + b(n-1)}$$

and

$$\xi = \frac{(bn-1)(n\sigma^2_{\text{ratio}}+1)}{(b-1)(n\sigma^2_{\text{ratio}}+1) + b(n-1)}$$

implying that the distribution of the F_{anova}-statistic and hence the type I

error rate is influenced by the relative magnitudes of σ_β^2 and σ^2.

The ratio σ_{ratio}^2 is a 1 to 1 function of the well-known intra-class correlation (ICC)

$$\text{ICC} = \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma^2}$$

or

$$\text{ICC} = \frac{1}{1 + \sigma_{\text{ratio}}^{-2}}$$

In the special case where $n = 1$ (one observation per group) then $\nu = a(b-1)$ and $\xi = 1$. In the case of $\sigma_\beta^2 = 0$ then $\nu = a(bn-1)$ and $\xi = 1$. In both cases the approximate distribution in (5) becomes the usual (appropriate) F-distribution and the type I error rate in (6) becomes α.

5 SIMULATION STUDY

A simulation study was conducted to verify the accuracy of the closed-form expressions for the type I error rate. Also, the impact on the type I error rate is assessed for different sample sizes and variance ratios when the hierarchical structure of data is disregarded.

Simulated type I error rates were obtained by simulating data from model (1). In all cases the between-group variation was fixed at $\sigma_\beta^2 = 0.1$ and the reported σ_{ratio}^2 was achieved by letting $\sigma^2 = \sigma_\beta^2/\sigma_{\text{ratio}}^2$. As the type I error rate is defined under H_0, data were generated with no treatment effect, i.e. $\tau_1, ..., \tau_a = 0$. For each combination (each cell in Table 1) 10000 simulations were conducted. The approximate type I error rates were calculated

from (6). Throughout, the nominal α was 0.05 unless stated otherwise. All simulations and calculations were performed using R, version 3.0.2 (R Core Team 2013). The `nlme` package (Pinheiro et al. 2013; Pinheiro and Bates 2000) was used to fit the linear mixed-effects model as defined in (1), and Wilson confidence intervals (Wilson 1927; Agresti and Coull 1998) of the simulated type I error rates were provided by the `binom` package (Dorai-Raj 2014).

The particular data structure addressed here (depicted in Figure 1) arises in many different scientific disciplines where the natural range of the parameters varies. For instance, animal experiments often imply small sample sizes and a sizeable variance ratio, σ^2_{ratio}, whereas studies comparing production methods on items from different batches may include a large number of batches and/or items per batch and the variance ratio may be somewhat smaller. The levels of the parameters, for which the type I error rates are obtained, are selected to reflect this diversity.

Table 1 summarizes the type I error rate for different combinations of treatments, groups per treatment, observations per group and ratios of the variance components. Increasing the number of treatments increases the type I error rate considerably. Similar conclusions can be made when the number of observations per group are increased. Interestingly, the number of groups per treatment does not affect the type I error rate noticeably. Increasing σ^2_{ratio} (increasing σ^2_β relative to σ^2) in general increases the type I error rate. For all cases type I error rate was greater than the nominal α at 0.05. Most

Table 1: Type I error rates for different combinations of number of treatments (a), groups per treatment (b), observations per group (n) and variance ratio $\sigma^2_{\text{ratio}} = \sigma^2_\beta/\sigma^2$. The approximate type I error rates were found from (5), and the simulated type I error rates were based on 10000 simulations for each combination (each cell).

Treatm.	Groups	Observ.	σ^2_{ratio}							
			0.25		0.50		1.00		2.00	
			Appr.	Sim.	Appr.	Sim.	Appr.	Sim.	Appr.	Sim.
2	2	2	0.076	0.074	0.099	0.095	0.134	0.127	0.177*	0.161
		10	**0.259**	**0.255**	**0.363**	**0.366**	**0.465**	**0.462**	**0.548**	**0.547**
		100	**0.683**	**0.682**	**0.760**	**0.765**	**0.813**	**0.816**	**0.849**	**0.847**
	50	2	0.074	0.074	0.090	0.087	0.110	0.108	0.130	0.131
		10	**0.242**	**0.240**	**0.328**	**0.332**	**0.406**	**0.405**	**0.462**	**0.457**
		100	**0.668**	**0.663**	**0.738**	**0.737**	**0.784**	**0.783**	**0.812**	**0.813**
6	2	2	0.107	0.107	0.160	0.161	**0.242**	**0.235**	**0.335**	**0.336**
		10	**0.594**	**0.590**	**0.788**	**0.794**	**0.899**	**0.901**	**0.948**	**0.948**
		100	**0.993**	**0.992**	**0.998**	**0.998**	**0.999**	**1.000**	**1.000**	**1.000**
	50	2	0.101	0.102	0.141	0.140	0.195	0.196	**0.251**	**0.249**
		10	**0.558**	**0.565**	**0.738**	**0.734**	**0.849**	**0.851**	**0.905**	**0.901**
		100	**0.991**	**0.990**	**0.997**	**0.998**	**0.999**	**0.999**	**0.999**	**1.000**

* Approximate type I error rates not covered by the 95% confidence intervals of the simulated type I error rates.
Type I error rates greater than 0.20 are marked in bold.

combinations gave type I error rates greater than 0.10 and more than half resulted in type I error rates greater than 0.50.

A visual representation of the inflated type I error rates is seen in Figure 2. This example shows the sampling distributions in case of six treatments, two groups per treatment, ten observations per group and with a variance ratio of $\sigma^2_{\text{ratio}} = 0.25$. The dashed line is the assumed sampling distribution of F_{anova} when the hierarchical structure is ignored, and the distribution from which the critical value is determined. This is not the actual sampling

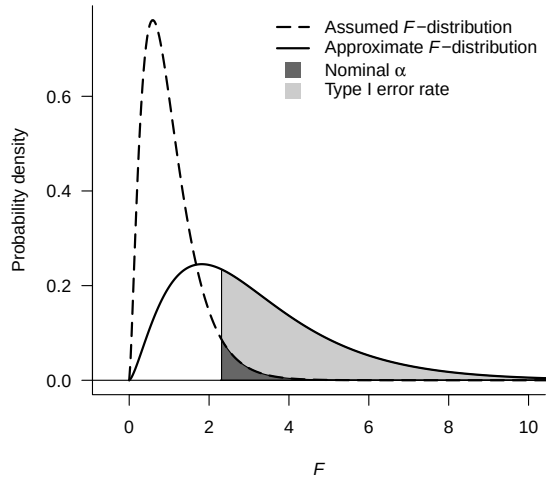

Figure 2: The F-distributions in case of six treatments, two groups per treatment, ten observations per group and $\sigma^2_{\text{ratio}} = 0.25$. The assumed F-distribution refers to the distribution from which the critical value is obtained. The approximate F-distribution is the distribution of F_{anova} as defined in (5). The approximate F-distribution has a heavier right tail implying that the type I error rate is greater than the nominal α at 0.05.

distribution, though, and the solid line depicts the approximate scaled F-distribution given in (5). The expected value is $E(F_{\text{anova}}) = 3.16$, which is greater than the usual expectation under H_0 at approximately 1. Additional skewness is thereby imposed on the F-distribution implying a heavier right tail. For the combinations shown in Table 1 the expectations $E(F_{\text{anova}})$ are between 1.21 and 134.56.

Proceeding with this example the type I error rate is shown as a function of the nominal α (black line) in Figure 3. The grey solid line indicates the

intended equality between the type I error rate and the nominal α. Both axes are on a logarithmic scale. It can be seen that irrespective of the level of the nominal α the type I error rate is considerably inflated when the hierarchical structure is ignored. Even when significance is demonstrated at a level of 0.001, which often is considered fairly strong evidence against H_0, it by no means guarantees that the actual type I error rate is anywhere near the conventional level of 0.05. In fact, for the example considered here a nominal α of 0.001 corresponds to a type I error rate of 0.220 while a nominal α of 0.01 corresponds to a type I error rate of 0.406.

The validity of the approximate type I error rates was assessed by making an informal comparison to the simulated type I error rates. Wilson 95% confidence intervals (CI) were obtained (not shown), and in all cases but one the approximate type I error rates were covered by the CI for the simulated type I error rates. This agrees with the expectation of a few values falling outside the CI given the number of comparisons and the confidence level. The one case not covered by the CI is marked with an asterisk in Table 1.

6 CONCLUSION

The goal of this study is to draw attention to the implications of ignoring the hierarchical structure of data during the analysis. We believe that this unfortunately is the case in various fields of research and application. The reason may be the failure to recognize the hierarchical structure or lack of ap-

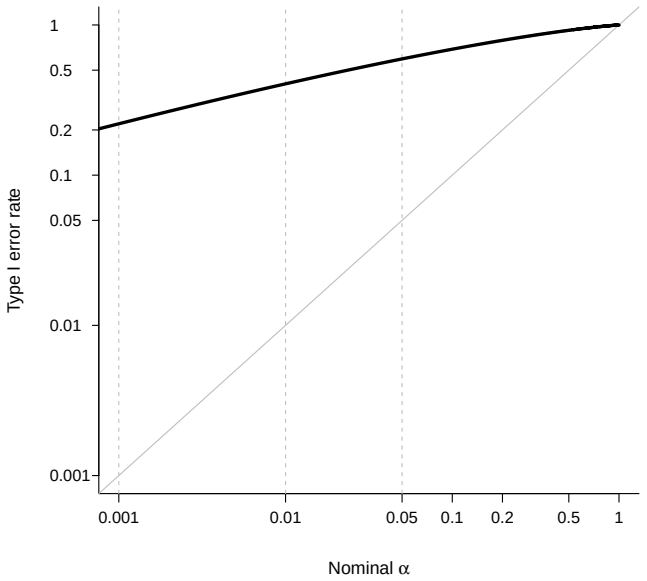

Figure 3: The type I error rate versus the nominal α (black line) in case of six treatments, two groups per treatment, ten observations per group and $\sigma^2_{\text{ratio}} = 0.25$. The grey solid reference line corresponds to equlity between the type I error rate and the nominal α. Both axes are on a logarithmic scale.

preciation of the severe consequences that are entailed when the hierarchical structure is disregarded.

For illustration purposes we focus on the simple yet widespread hierarchical structure shown in Figure 1. We believe that similar and potentially more striking results can be expected when data exhibit more complicated hierarchical structures. To clarify this further an even more detailed study is required. However this is beyond the scope of this study.

When the hierarchical structure is ignored, the sampling distribution is

no longer the usual non-scaled F-distribution, but instead it is approximately a scaled F-distribution. This imposes additional skewness to the sampling distribution as opposed to when all observations are independent. Hence, the more extreme values of the observed statistic are inferred to be significant more often than they should, and this leads to inflation of the type I error rate.

For different parameter values seemingly within a realistic range (depending on the particular field), the inflation of the type I error rate is rather substantial. Thus, situations will occur where it becomes extremely likely that the treatment effect is inferred to be significant, when in fact it is not, and incorrect conclusions are thereby drawn.

It is of concern that studies with a positive finding, i.e. a significant treatment effect, are more likely to be published (Easterbrook et al. 1991; Song et al. 2010). The results from the current study confirm that null results from studies that fail to account for the hierarchical structure are more likely to appear significant and thus are more likely to be published. It is therefore possible that these spurious findings are over-represented in the literature, so that evidence against the hypothesis of the treatments being ineffective fallaciously appears to accumulate.

REFERENCES

Agresti, A. and Coull, B. A. (1998), "Approximate is Better than "Exact" for Interval Estimation of Binomial Proportions," *The American Statistician*, 52, 119–126.

Almeida, M. R., Aissa, A. F., Gomes, T. D. U. H., Darin, J. D. C., Chisté, R. C., Mercadante, A. Z., Antunes, L. M. G., and Bianchi, M. L. P. (2013), "*In Vivo* Genotoxicity and Oxidative Stress Evaluation of an Ethanolic Extract from Piquiá (*Caryocar villosum* Pulp)," *Journal of Medicinal Food*, 16, 268–271.

Baccaglini, L., Shuster, J. J., Cheng, J., Theriaque, D. W., Schoenbach, V. J., Tomar, S. L., and Poole, C. (2010), "Design and statistical analysis of oral medicine studies: common pitfalls," *Oral Diseases*, 16, 233–241.

Baldwin, S. A., Murray, D. M., and Shadish, W. R. (2005), "Empirically Supported Treatments or Type I Errors? Problems With the Analysis of Data From Group-Administered Treatments," *Journal of Consulting and Clinical Psychology*, 73, 924–935.

Box, G. E. P. (1954), "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification," *The Annals of Mathematical Statistics*, 25, 290–302.

Chekhun, V. F., Lozovskaya, Y. V., Lukyanova, Y., Demash, D. V., Todor, I. N., and Naleskina, L. A. (2013), "Assessment of the Cyto- and Genotoxic Effects of a Nanoferromagnetic and a Static Magnetic Field In Vivo," *Cytology and Genetics*, 47, 179–187.

Dorai-Raj, S. (2014), *binom: Binomial Confidence Intervals For Several Parameterizations*, r package version 1.1-1.

Easterbrook, P. J., Berlin, J. A., Gopalan, R., and Matthews, D. R. (1991), "Publication bias in clinical research," *Lancet*, 337, 867–872.

Freedman, D. A. (2001), "Ecological inference and the ecological fallacy," in *International Encyclopedia of the Social and Behavioral Sciences*, eds. Smelser, N. J. and baltes, P. B., Elsevier, vol. 6, pp. 4027–4030.

Gelman, A. and Hill, J. (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models*, New York: Cambridge University Press.

Gelman, A., Park, D., Shor, B., Bafumi, J., and Cortina, J. (2008), *Red State, Blue State, Rich State, Poor State*, Princeton University Press.

Hansen, M. K. and Kulahci, M. (2014), "The type I error rate for *in vivo* Comet assay data when the hierarchical structure is disregarded," Tech. Rep. 9, Department of Applied Mathematics and Computer Science, Technical University of Denmark.

Holson, R. R. and Pearce, B. (1992), "Principles and Pitfalls in the Analysis

of Prenatal Treatment Effects in Multiparous Species," *Neurotoxicology and Teratology*, 14, 221–228.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995), *Continuous univariate distributions*, vol. 2 of *Wiley series in probability and mathematical statistics*, John Wiley & Sons, 2nd ed.

Kelly, A. D. R., Lemaire, M., Young, Y. K., Eustache, J. H., Guilbert, C., Molina, M. F., and Mann, K. K. (2013), "*In Vivo* Tungsten Exposure Alters B-Cell Development and Increases DNA Damage in Murine Bone Marrow," *Toxicological Sciences*, 131, 434–446.

Kenny, D. A. and Judd, C. M. (1986), "Consequences of Violating the Independence Assumption in Analysis of Variance," *Psychological Bulletin*, 99, 422–431.

Kromrey, J. D. and Dickinson, W. B. (1996), "Detecting unit of analysis problems in nested designs: statistical power and type I error rates of the F test for groups-within-treatments effects," *Educational and Psychological Measurement*, 56, 215–231.

Laird, N. M. and Ware, J. H. (1982), "Random-Effects Models for Longitudinal Data," *Biometrics*, 38, 963–974.

McCoach, D. B. and Adelson, J. L. (2010), "Dealing With Dependence (Part I): Understanding the Effects of Clustered Data," *Gifted Child Quarterly*, 54, 152–155.

Montgomery, D. C. (2005), *Design and Analysis of Experiments*, USA: John Wiley & Sons, Inc., sixth ed.

Morales, M., Martínez-Paz, P., Ozáez, I., Martínez-Guitarte, J. L., and Morcillo, G. (2013), "DNA damage and transcriptional changes induced by tributyltin (TBT) after short *in vivo* exposures of *Chironomus riparius* (Diptera) larvae," *Comparative Biochemistry and Physiology, Part C*, 158, 57–63.

Musca, S. C., Kamiejski, R., Nugier, A., Méot, A., Er-Rafiy, A., and Brauer, M. (2011), "Data with hierarchical structure: impact of intraclass correlation and sample size on Type-I error," *Frontiers in Psychology*, 2.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2013), *nlme: Linear and Nonlinear Mixed Effects Models*, r package version 3.1-113.

Pinheiro, J. C. and Bates, D. M. (2000), *Mixed-Effects Models in S and S-PLUS*, Statistics and Computing Series, New York, NY: Springer-Verlag.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Raudenbush, S. W. and Bryk, A. S. (2002), *Hierarchical Linear Models*, Thousand Oaks, USA: Sage Publications, 2nd ed.

Robinson, W. S. (1950), "Ecological correlations and the behavior of individuals," *American Sociological Review*, 15, 351–357.

Satterthwaite, F. E. (1941), "Synthesis of Variance," *Psychometrika*, 6, 309–316.

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., Hing, C., Kwok, C. S., Pang, C., and Harvey, I. (2010), "Dissemination and publication of research findings: an updated review of related biases," *Health Technology Assessment*, 14, 1–193.

Tasca, G. A., Illing, V., Joyce, A. S., and Ogrodniczuk, J. S. (2009), "Three-level multilevel growth models for nested change data: A guide for group treatment researchers," *Psychotherapy Research*, 19, 453–461.

Verbeke, G. and Molenberghs, G. (2009), *Linear Mixed Models for Longitudinal Data*, Springer Series in Statistics, New York, NY, USA: Springer.

Wampold, B. E. and Serlin, R. C. (2000), "The Consequence of Ignoring a Nested Factor on Measures of Effect Size in Analysis of Variance," *Psychological Methods*, 5, 425–433.

Welch, B. L. (1938), "The Significance of the Difference Between Two Means when the Population Variances are Unequal," *Biometrika*, 29, 350–362.

Wilson, E. B. (1927), "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, 22, 209–212.

Appendix D

In vivo Comet assay - statistical analysis and power calculations of mice testicular cells

Hansen, M. K., Sharma, A. K., Dybdahl, M., Boberg, J. and Kulahci, M. (2014). *In vivo* Comet assay - statistical analysis and power calculations of mice testicular cells. *Mutation Research - Genetic Toxicoloty and Environmental Mutagenesis*, *774*, 29-40. 10.1016/j.mrgentox.2014.08.006.

Mutation Research 774 (2014) 29–40

Contents lists available at ScienceDirect

Mutation Research/Genetic Toxicology and Environmental Mutagenesis

journal homepage: www.elsevier.com/locate/gentox

Community address: www.elsevier.com/locate/mutres

In vivo Comet assay – statistical analysis and power calculations of mice testicular cells

Merete Kjær Hansen[a,*], Anoop Kumar Sharma[b], Marianne Dybdahl[b], Julie Boberg[b], Murat Kulahci[a,c]

[a] Technical University of Denmark, Department of Applied Mathematics and Computer Science, Matematiktorvet, DK-2800 Kgs. Lyngby, Denmark
[b] Technical University of Denmark, National Food Institute, Mørkhøj Bygade 19, DK-2860 Søborg, Denmark
[c] Luleå University of Technology, Department of Business Administration, Technology and Social Sciences, Luleå, Sweden

ARTICLE INFO

Article history:
Received 30 January 2014
Received in revised form 16 August 2014
Accepted 29 August 2014
Available online 8 September 2014

Keywords:
Single-cell gel electrophoresis
Genotoxicity
DNA damage
Germ cells
Power
Statistical analysis

ABSTRACT

The *in vivo* Comet assay is a sensitive method for evaluating DNA damage. A recurrent concern is how to analyze the data appropriately and efficiently. A popular approach is to summarize the raw data into a summary statistic prior to the statistical analysis. However, consensus on which summary statistic to use has yet to be reached. Another important consideration concerns the assessment of proper sample sizes in the design of Comet assay studies. This study aims to identify a statistic suitably summarizing the % tail DNA of mice testicular samples in Comet assay studies. A second aim is to provide curves for this statistic outlining the number of animals and gels to use. The current study was based on 11 compounds administered *via* oral gavage in three doses to male mice: CAS no. 110-26-9, CAS no. 512-56-1, CAS no. 111873-33-7, CAS no. 79-94-7, CAS no. 115-96-8, CAS no. 598-55-0, CAS no. 636-97-5, CAS no. 85-28-9, CAS no. 13674-87-8, CAS no. 43100-38-5 and CAS no. 60965-26-6. Testicular cells were examined using the alkaline version of the Comet assay and the DNA damage was quantified as % tail DNA using a fully automatic scoring system. From the raw data 23 summary statistics were examined. A linear mixed-effects model was fitted to the summarized data and the estimated variance components were used to generate power curves as a function of sample size. The statistic that most appropriately summarized the within-sample distributions was the median of the log-transformed data, as it most consistently conformed to the assumptions of the statistical model. Power curves for 1.5-, 2-, and 2.5-fold changes of the highest dose group compared to the control group when 50 and 100 cells were scored per gel are provided to aid in the design of future Comet assay studies on testicular cells.

1. Introduction

The Comet assay (also known as the single cell gel electrophoresis assay) is a simple, rapid, and sensitive technique for measuring DNA strand breaks within individual mammalian cells. The *in vivo* Comet assay is emerging as the default second *in vivo* assay to follow-up *in vitro* positive genotoxicity results for regulatory and mechanistic purposes. Typically, target organs of toxicological relevance or site of contact are selected for analysis [1].

Statistical issues in the Comet assay have been addressed in the last decade and a thorough review of the statistical aspects of the Comet assay is given by Lovell and Omori [2], addressing considerations both with respect to the design of the study and the statistical analysis. A recurrent concern is how the Comet assay data are analyzed appropriately and efficiently. The collection of models commonly referred to as ANOVA all assume normally distributed data. The asymmetric distribution obtained from each sample therefore impede a direct application of standard ANOVA methods and more advanced analysis strategies must be considered [3]. Alternatively, a summary statistic representing each within-sample distribution may be extracted and subsequently subjected to a proper analysis. The latter approach is commonly practiced although no consensus as to which summary statistic to employ has been reached. Also, it remains unclear if a logarithmic transformation prior to the summary calculation is requisite [2]. Commonly used summary statistics calculated from raw or log-transformed data include the mean [4–6] and median [4–7] but other statistics such as the 75th [5,7] and the 90th [5,6] percentile have also been suggested. However, this issue has not been thoroughly examined and the few studies specifically addressing this topic also concern the tail length and tail moment [6,7]. Which summary statistic to use for the % tail DNA has not been investigated comprehensively.

* Corresponding author. Tel.: +45 4525 5351; fax: +45 4588 1399.
E-mail address: mkha@dtu.dk (M.K. Hansen).

http://dx.doi.org/10.1016/j.mrgentox.2014.08.006
1383-5718/

Often Comet assay data are analyzed by means of a one-way ANOVA with dose as a fixed effect. Thereby, another factor, namely the animal, seems to be inadvertently ignored in many studies. Observations from the same animal are not independent and ignoring this factor in the model when it indeed induces variation in the observed data thus violates the most critical assumption of the statistical model. Since the animals are selected at random, this second factor should be added to the model as a random factor. The presence of both fixed effects and random effects is exactly what characterizes a mixed-effects model. The nature of the summarized data therefore implies that a suitable analysis could be conveyed by means of a linear mixed-effects model with dose as a fixed effect and animal as a random effect. A natural consequence of this modelling approach is that observations from the same animal are allowed to be more similar than observations obtained from different animals. It thus relaxes the most important assumption made by the fixed-effects model, namely the assumption of independence. Still, the mixed-effects model makes a number of assumptions that must be met in order to ensure proper inference, that is, to avoid inflation of the rate of false negatives or false positives. Furthermore, it should be taken into account that the uncertainty of the estimated summary statistic vary considerably, hence some estimates are more reliably determined than others.

One important but sometimes overlooked concern in planning a particular study is the determination of an appropriate sample size in order to achieve adequate statistical power. Power is defined as the probability of correctly rejecting the null hypothesis when it is indeed false and, although influenced by several factors, power in general increases with increasing sample size when all other things are held constant. The power and sample size consideration is treated in a simulation study by Wiklund and Agurell [6] for the two end points tail length and tail moment on data from mice white blood cells and different mice liver cells. For both tail length and tail moment the mean, median and 90th percentile of the raw data and of the log-transformed data are evaluated. Sample size recommendations for the end point % tail DNA are provided by Smith et al. [8], who used the mean of the log-transformed data as a summary statistic from samples of rat liver, blood, bone marrow and stomach samples.

Because of reasons such as reduced seamen quality and reduced fertility in the Western world, there is a growing interest in the evaluation of genotoxicity in male germ cells [9]. The comet assay has the potential to detect germ cell genotoxicity and may be used for demonstrating the ability of a substance or its metabolite(s) to directly interact with the genetic material and causing DNA damage of gonadal and/or germ cells [10]. Genotoxicity data of gonadal and/or germ cells are used in hazard assessment and in the classification/labelling of substances. Because of the growing interest in assessing the genotoxicity of male gonadal and/or germ cells together with the lack of analysis of summary statistics in comet assay data of testicular cells, the focus is on testicular cells in this study.

The present study is part of a project with the overall purpose of extending and improving an existing quantitative structure-activity relationship (QSAR) computer model for *in vivo* Comet assay, developed at the National Food Institute, Technical University of Denmark. The aim of this paper is 2-fold. First, a range of candidate summary statistics extracted from several studies were evaluated with respect to some established criteria in order to recommend the 'best' summary statistic for testicular cell data. Second, power curves for testicular cell data illustrating power as a function of sample size were generated for this particular summary statistic in order to facilitate the choice of animals employed per group and the number of gels to use per animal.

2. Materials and methods

2.1. Chemicals tested

The chemicals were strategically selected in order to improve an existing QSAR model for *in vivo* Comet assay. This was done by selecting chemical groups not already covered by the model or chemicals that could strengthen the predictive statistics of specific structural fragments. The QSAR model is based on mice data; therefore, mice were used in this study. 11 chemicals were tested with CAS numbers: 110-26-9 (Acrylamide, *N*,*N*′-methylenebis-), 512-56-1 (Methyl phosphate phosphoric acid, trimethyl ester), 111873-33-7 (Perfluorooctane sulfonate), 79-94-7 (Tetrabromobisphenol A), 115-96-8 (Tris(2-chlorethyl) phosphate) and 598-55-0 (Methylcarbamate), 636-97-5 (4-Nitrobenzoic hydrazide), 85-28-9, (4-chloro-2-hydroxy-4-methoxybenzophenone), 13674-87-8 (Tris[2-chloro-1-(chloromethyl)ethyl]phosphate), 43100-38-5 (4-*tert*-Butylbenzoic hydrazide A), 60965-26-6 (2-Bromo-2′,4′-dimethoxyacetophenone). Ethyl methanesulfonate (EMS, CAS no. 62-50-0) was used as the positive control. All the chemicals were obtained from Sigma–Aldrich, Brondby, Denmark. To our knowledge, limited information is available in the literature about the genotoxicity of the chemicals, apart from CAS number 110-26-9. Table 1 summarizes the published genotoxicity results of the chemicals.

2.2. Animals and animal husbandry

CAS no. 110-26-9 and CAS no. 512-56-1 were tested in one animal study (study 1) and CAS no. 111873-33-7, 79-94-7, 115-96-8 and 598-55-0 were tested in another animal study (study 2). CAS no. 636-97-5, 85-28-9, 13674-87-8, 43100-38-5 and 60965-26-6 were tested in a third animal study (study 3). In total 205 CD-1 male mice, 4 weeks of age were purchased from Taconic MB, DK-4623, Lille Skensved, Denmark. Animals were allowed to acclimatize for a week. The weight (mean ± SD) of the 40 mice at arrival in study 1: 29.8 ± 1.2 g, 75 mice in study 2: 30.5 ± 1 g and 90 mice in study 3: 33.6 ± 2.3 g. The mice were randomly divided into dose groups and were housed individually in cages (Macrolon type III high, Techniplast Gazzada S ar. L., Buguggiate, Italy) with wood bedding (Tapvei, Finland) under controlled environmental conditions (temperature 22 ± 1 °C, relative humidity 55 ± 5%, 12 h light/dark cycle, air changed 10 times/h) and had free access to feed (Altromin 1324, Lage, Germany) and tap water acidified with citric acid, pH = 3.5 (to prevent growth of microorganisms). During the acclimatization and study periods all mice were observed at least twice daily for any abnormalities in clinical appearance.

Table 1
Published *in vivo* genotoxicity results of the tested chemicals.

CAS number	Reference	Published *in vivo* comet assay data of gonadal/sperm cells	Published *in vivo* genotoxicological data
110-26-9	[26–28]	Positive response in mice gonadal sperm cells and testicular somatic cells. Positive response in rat testicular somatic cells. Negative in rat gonadal sperm cells.	Positive response in comet assay in mice blood, liver, and duodenum. Positive response in comet assay in rat blood, duodenum and thyroid. Negative in rat liver. Positive response in *in vivo* micronucleus assay (mice) in peripheral blood and bone marrow (male rats and mice). Positive response in other assays *e.g.* DNA adducts
111873-33-7	[29]		Positive response in bone marrow cells (female rats) in comet assay and micronucleus assay.
115-96-8	[30,31]		Positive response in Chinese hamster bone marrow cells in the micronucleus test (two males and two females per dose group). Negative in the micronucleus test (mice).
512-56-1	[32]		Positive response in the dominant lethal mutation assay.
598-55-0	[34]		Negative in sex linked mutations in Drosophila.
13674-87-8	[35]		Negative in induction of chromosomal aberrations in rats. Negative in induction of clastogenic effects in mice.

2.3. Experimental design

The doses were chosen for the 11 chemicals according to the published literature of *in vivo* genotoxicity test data and experimental LD_{50} values after oral exposure of mice. If no experimental data were available, a predicted QSAR LD_{50} value was used for estimating the test doses (Pharma ToxBoxes version 1, now ACD/Labs). The maximum dose was about 50% of the LD_{50} values, however no doses were tested above the maximum recommendation of 2000 mg/kg bw. The following doses were administered; CAS no. 110-26-9: 50, 100 and 190 mg/kg bw, CAS no. 512-56-1: 125, 250 and 500 mg/kg bw, CAS no. 111873-33-7: 100, 200 and 300 mg/kg bw, CAS no. 79-94-7: 500, 1000 and 2000 mg/kg bw, CAS no. 115-96-8: 500, 1000 and 1500 mg/kg bw, CAS no. 598-55-0: 500, 1000 and 2000 mg/kg bw, CAS no. 85-28-9: 500, 1000 and 1500 mg/kg bw, CAS no. 13674-87-8: 225, 450 and 900 mg/kg bw, CAS no. 43100-38-5: 62.5, 125 and 250 mg/kg bw, CAS no. 60965-26-6: 45, 90 and 180 mg/kg bw and CAS no. 636-97-5: 12.5, 25 and 50 mg/kg bw. Each dose group consisted of five mice.

For CAS no. 110-26-9, 512-56-1, 598-55-0 and 636-97-5 water was used as the solvent and for CAS no. 111873-33-7, 79-94-7, 115-96-8, 60965-26-6, 43100-38-5, 13674-87-8 and 85-28-9 corn oil was used. A positive control group of five mice administered to 300 mg/kg bw EMS (water as solvent) was included in each of the three animal studies. Control groups of five mice administered to solvent were also included; water in study 1, water and corn oil in study 2 and 3. The mice were dosed orally by gavage twice 24 h apart. Dosing suspensions were freshly prepared prior to each dosing occasion and given in a volume of 1 ml/100 g bw. All doses were placed on a magnetic stirrer until dosing. All animals were fasted overnight before the first dosing. Two to four hours after the second dosing the animals were anaesthetized in CO_2/O_2 and decapitated. After macroscopic examination, the testicles were excised and weighed. After removing the capsule the right testicles were put in cryotubes and used later in the Comet assay. Freezing medium was not added because it was an organ that was frozen and not cell suspensions. The cryotubes were immediately frozen in Mr. Frosty for about 1 h and then transferred to −80 °C freezer until analyzed in the Comet assay. The left testicles were fixed in Bouin's fixative and routinely processed for paraffin fixation. One section (3 μM) per testis were evaluated by an experienced pathologist. A detailed qualitative examination of the testes was made, taking into account the tubular stages of the spermatogenic cycle. The examination was conducted to identify treatment-related effects such as missing germ cell layers or types, retained spermatids, multinucleate or apoptotic germ cells and sloughing of spermatogenic cells into the lumen.

A testicle comprises different cell populations including somatic supportive cells called Sertoli cells, differentiating germ cells in various stages of spermatogenesis and spermiogenesis, interstitial Leydig cells, macrophages, and fibroblasts as well as blood vessels and lymphatic vessels containing different cell populations. Therefore, the DNA isolated from the testicular tissue origins from a mixture of different cell types. The animal study was performed under conditions approved by The Danish Agency of Protection of Experimental Animals and the in-house Animal Welfare Committee.

2.4. Alkaline Comet assay

The Comet assay was performed according to [11] following the recommendations of [12], with some minor modifications according to the manufacturer of the CometAssay® Kit (Trevigen, Gaithersburg, Maryland). The cryotubes with the testicules from each mouse were added with 1.5 ml ice cold mincing solution (Hank's balanced salt solution (Ca^{2+}, Mg^{2+} free) with 20 mM EDTA and 10% dimethyl sulfoxide). The testicules were gently crushed 4–5 times with a pastil. The solution was filtered through a 100 μm nylon filter (BD Falcon, Sigma–Aldrich, Denmark). Then, the solution was centrifuged at 1200 rpm for 5 min at 4 °C (Eppendorf Centrifuge 5810R, Buch & Holm, Herlev, Denmark). The supernatant was removed and the pellet was resuspended in 1.5 ml mincing solution. This solution was filtered through a 100 μm nylon filter. Three microlitres of this suspension was mixed with 150 μl of the molten CometAssay TM LMAgarose (Trevigen, Gaithersburg, Maryland, US). Thirty microlitres of this mixture was applied onto one sample area of two gels on two different slides (one gel on one slide) consisting of 20 gels (CometSlide™ HT, Trevigen, Gaithersburg, Maryland, US). After solidification the embedded cells were lysed in a cold alkaline lysis buffer for 60 min. For DNA unwinding the slides were placed in the alkaline electrophoresis solution (pH > 13) in the electrophoresis jar at 4 °C for 40 min, and electrophoresis was run in the same buffer for 30 min at 4 °C (1 V/cm and 270 mA). After neutralization, fixation in 96% ethanol and DNA staining with 10 μl SYBR Green on all gels of the slides and a drop of antifade solution was added to each gel to avoid fading. Negative (Caco-2 cells in culture medium) and positive (Caco-2 cells exposed to 200 and 400 μM ethyl methanesulfonate for 30 min) controls were included with each 20 gels slide for each electrophoresis run. Testicule samples from each mouse were analyzed in two gels on two different slides.

Fully automatic Comet assay scoring was performed using the Pathfinder™ Cellscan Comet imaging system (IMSTAR, Paris, France). The system is described in details in [13]. Tail intensity (% tail DNA) of each comet was used. The number of cells scored on the gels depended on the cell density of the gels. In this study 100 random cells were subsequently selected from each gel unless it was explicitly stated otherwise, hence for each mouse 200 cells were scored (100 cells per gel and two gels per mouse). For three of the power curves shown in Fig. 3 (first column) 50 cells were sampled per gel (100 cells per mouse). For the three remaining power curves (second column) 100 cells were sampled per gel (200 cells per mouse). Selection of the 50 and 100 random cells for each gel was done using the `sample` function in R, version 3.0.2 [14]. Cells with high levels of DNA damage *i.e.* hedgehogs were not discarded. The number of highly damaged cells with % tail DNA >80% for each gel were recorded.

2.5. Statistical analysis

Statistical analysis: For each gel the following summary statistics were calculated: the mean, median (50th), 55th, 60th, 65th, 70th, 75th, 80th, 85th, 90th and 95th percentile. The % tail DNA measured for each cell is naturally restricted to be non-negative and the distribution of the % tail DNA within each sample are strongly positively skewed. A popular endeavour to normalize the data and/or stabilize the variance of such data is to take the natural logarithm. Each of the statistics was thus calculated from the raw data as well as from data subjected to the natural logarithm. As some observations were recorded as zero, a small constant (0.001) was added to data when calculating the mean of the log-transformed data to avoid taken the logarithm of zero. Additionally, one summary statistic was calculated as the mean of the raw data and subsequently log-transformed. This measure will be referred to as the log(mean). In total, 23 candidate summary statistics were extracted.

The summarized data was fitted using a linear mixed-effects model with dose as a fixed effect and animal as a random effect. The animals in one dose group are different from the animals in other dose groups, and this induces a nested structure in data as illustrated in Figure 1. The linear mixed-effects model with animal nested within dose is:

$$y_{ijk} = \mu + d_i + A_{(i)j} + \varepsilon_{(ij)k} \quad (1)$$

where

$$i = 1, \ldots, 4, \quad j = 1, \ldots, 5, \quad k = 1, \ldots, 2,$$

$$A_{(i)j} \sim N(0, \sigma_A^2), \quad \varepsilon_{(ij)k} \sim N(0, \sigma^2).$$

y_{ijk} is the summary statistic of interest calculated for each gel and μ and d_i are the fixed effects for the intercept and dose, respectively. $A_{(i)j}$ is the random effect of the jth animal nested within the ith dose and $\varepsilon_{(ij)k}$ is the within-group error. The parentheses in the subscripts indicate the nesting structure with the parent level(s) given inside the parentheses. See Montgomery [15] for a more elaborate exposition of the linear mixed-effects model with nested effects. The criteria for selecting a summary statistic and all power calculations were based on this modelling approach. Dunnett's test was subsequently applied to compare the three dose groups to the corresponding control group. Example of R code for fitting model (1) to Comet assay data is seen in appendix A. The values given in Table 2 were calculated by first averaging the two summary statistics obtained for each animal (*i.e.* one summary statistics for each gel and two gels per animal) and from these values the average and standard deviation were calculated.

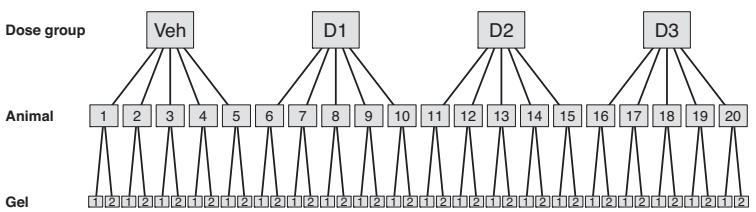

Fig. 1. Outline of the design of the conducted *in vivo* Comet assay studies. Animals in one dose group are different from animals in the other dose groups, and this imposes a nested structure of the study design and the resulting data. Veh, D1, D2, D3 indicate a vehicle group and three dose groups, respectively.

Table 2a
Obtained % tail DNA after administration of CAS no. 110-26-9 (Acrylamide, *N,N'*-methylenebis-) and CAS no. 512-56-1 (methyl phosphate phosphoric acid, trimethyl ester) in study 1. Average % tail DNA with SD in parentheses. The values were calculated by first averaging the two summary statistics for each animal and from these values the average and SD were calculated. There were five animals in each dose group and 200 cells were scored for each mouse. Data were analyzed by means of a linear mixed-effects model as defined in model (1) with Dunnett's test to compare the dose groups to their corresponding control. Values in bold indicate a significant difference. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

CAS no.	0 mg/kg bw		50 mg/kg bw		100 mg/kg bw		190 mg/kg bw		Positive control (EMS)	
110-26-9	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)
Mean	7.1 (1.9)	1.0 (0.2)	6.7 (2.9)	0.8 (0.5)	8.3 (0.9)	1.3 (0.1)	**14.0 (3.7)*****	**1.7 (0.2)*****	**12.4 (1.8)*****	**1.9 (0.2)*****
log(mean)	1.9 (0.3)	–	1.8 (0.4)	–	2.1 (0.1)	–	**2.6 (0.3)*****	–	**2.5 (0.2)*****	–
Median	2.5 (0.6)	0.8 (0.3)	2.2 (1.8)	0.6 (0.7)	**4.4 (0.4)***	**1.5 (0.1)****	**7.8 (1.5)*****	**2.0 (0.2)*****	**8.7 (2.6)*****	**2.1 (0.3)*****
65th perc.	4.7 (2.1)	1.4 (0.4)	5.9 (3.0)	1.7 (0.6)	**8.0 (0.8)***	**2.1 (0.1)****	**12.5 (2.8)*****	**2.5 (0.2)*****	**12.5 (2.6)*****	**2.5 (0.2)*****
75th perc.	9.5 (4.7)	2.1 (0.4)	9.7 (4.2)	2.2 (0.4)	12.0 (2.1)	2.5 (0.2)	**19.1 (6.3)****	**2.9 (0.4)****	**16.0 (2.4)****	**2.8 (0.2)****
85th perc.	14.9 (5.0)	2.6 (0.3)	14.5 (6.5)	2.6 (0.5)	16.4 (2.8)	2.7 (0.2)	**27.8 (11.6)***	**3.2 (0.4)***	**21.3 (3.4)***	**3.0 (0.2)****
95th perc.	25.4 (3.9)	3.2 (0.2)	25.0 (12.2)	3.1 (0.5)	28.2 (4.1)	3.3 (0.3)	**49.4 (19.1)****	**3.8 (0.5)****	**38.1 (3.9)*****	**3.6 (0.1)*****

CAS no.	0 mg/kg bw		125 mg/kg bw		250 mg/kg bw		500 mg/kg bw		Positive control (EMS)	
512-56-1	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)
Mean	7.1 (1.9)	1.0 (0.2)	7.9 (3.7)	0.9 (0.4)	9.8 (2.2)	1.4 (0.1)	**13.9 (3.6)*****	**1.9 (0.4)*****	**12.4 (1.8)*****	**1.9 (0.2)*****
log(mean)	1.9 (0.3)	–	2.0 (0.5)	–	2.3 (0.2)	–	**2.6 (0.3)****	–	**2.5 (0.2)*****	–
Median	2.5 (0.6)	0.8 (0.3)	2.1 (0.9)	0.6 (0.5)	4.4 (0.4)	**1.5 (0.1)****	**8.2 (3.2)*****	**2.0 (0.4)*****	**8.7 (2.6)*****	**2.1 (0.3)*****
65th perc.	4.7 (2.1)	1.4 (0.4)	5.4 (3.1)	1.5 (0.6)	8.0 (0.6)	**2.1 (0.1)***	**12.4 (4.4)*****	**2.5 (0.4)*****	**12.5 (2.6)*****	**2.5 (0.2)*****
75th perc.	9.5 (4.7)	2.1 (0.4)	10.8 (7.0)	2.2 (0.7)	11.7 (3.0)	2.4 (0.3)	**17.9 (6.8)***	**2.8 (0.4)***	**16.0 (2.4)****	**2.8 (0.2)****
85th perc.	14.9 (5.0)	2.6 (0.3)	16.6 (8.9)	2.7 (0.5)	21.6 (11.0)	3.0 (0.5)	**28.8 (7.3)***	**3.3 (0.3)***	**21.3 (3.4)***	**3.0 (0.2)****
95th perc.	25.4 (3.9)	3.2 (0.2)	32.7 (15.6)	3.4 (0.5)	38.7 (11.3)	3.6 (0.3)	**46.6 (8.5)****	**3.8 (0.2)****	**38.1 (3.9)*****	**3.6 (0.1)*****

Table 2b
Obtained % tail DNA after administration of CAS no. 111873-33-7 (Perfluorooctane sulfonate) and CAS no. 79-94-7 (Tetrabromobisphenol A) in study 2. Average % tail DNA with SD in parentheses. The values were calculated by first averaging the two summary statistics for each animal and from these values the average and SD were calculated. There were five animals in each dose group and 200 cells were scored for each mouse. Data were analyzed by means of a linear mixed-effects model as defined in model (1) with Dunnett's test to compare the dose groups to their corresponding control. Values in bold indicate a significant difference. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

CAS no.	0 mg/kg bw		100 mg/kg bw		200 mg/kg bw		300 mg/kg bw		Positive control (EMS)	
111873-33-7	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)
Mean	6.1 (2.9)	0.7 (0.6)	**9.8 (1.0)****	1.0 (0.2)	8.1 (1.1)	0.8 (0.4)	6.5 (2.0)	0.8 (0.4)	**12.2 (1.8)*****	**1.7 (0.3)****
log(mean)	1.7 (0.5)	–	**2.3 (0.1)****	–	2.1 (0.2)	–	1.8 (0.3)	–	**2.5 (0.1)*****	–
Median	2.7 (1.7)	0.9 (0.6)	2.7 (0.8)	1.0 (0.3)	2.4 (0.9)	0.8 (0.4)	2.5 (1.1)	0.8 (0.3)	**7.6 (1.7)*****	**2.0 (0.2)*****
65th perc.	4.4 (2.6)	1.3 (0.6)	5.6 (1.2)	1.7 (0.2)	4.0 (1.5)	1.3 (0.4)	4.3 (2.1)	1.4 (0.4)	**10.8 (2.2)*****	**2.4 (0.2)*****
75th perc.	6.4 (4.0)	1.7 (0.6)	10.2 (2.7)	2.3 (0.2)	6.7 (2.2)	1.8 (0.3)	6.4 (3.3)	1.7 (0.5)	**13.9 (2.4)*****	**2.6 (0.2)****
85th perc.	9.5 (5.5)	2.1 (0.6)	**19.5 (3.7)****	**2.9 (0.2)****	14.0 (3.0)	2.6 (0.2)	11.7 (5.5)	2.4 (0.5)	**19.2 (3.4)*****	**2.9 (0.2)****
95th perc.	24.3 (13.0)	3.1 (0.5)	**46.7 (4.8)*****	**3.8 (0.1)*****	**43.7 (5.0)*****	**3.7 (0.1)****	26.9 (7.1)	3.2 (0.3)	**42.9 (4.9)****	**3.7 (0.1)****

CAS no.	0 mg/kg bw		500 mg/kg bw		1000 mg/kg bw		2000 mg/kg bw		Positive control (EMS)	
79-94-7	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)
Mean	6.1 (2.9)	0.7 (0.6)	8.3 (1.5)	0.8 (0.4)	8.9 (2.6)	0.8 (0.5)	7.0 (1.0)	0.5 (0.2)	**12.2 (1.8)*****	**1.7 (0.3)****
log(mean)	1.7 (0.5)	–	2.1 (0.2)	–	2.1 (0.3)	–	1.9 (0.2)	–	**2.5 (0.1)*****	–
Median	2.7 (1.7)	0.9 (0.6)	2.5 (1.3)	0.8 (0.5)	2.6 (1.6)	0.8 (0.5)	1.9 (0.5)	0.5 (0.2)	**7.6 (1.7)*****	**2.0 (0.2)*****
65th perc.	4.4 (2.6)	1.3 (0.6)	5.5 (2.4)	1.6 (0.5)	4.9 (2.1)	1.5 (0.4)	3.4 (0.9)	1.1 (0.3)	**10.8 (2.2)*****	**2.4 (0.2)*****
75th perc.	6.4 (4.0)	1.7 (0.6)	8.5 (3.2)	2.1 (0.4)	8.7 (4.1)	2.1 (0.5)	5.9 (2.0)	1.7 (0.4)	**13.9 (2.4)*****	**2.6 (0.2)****
85th perc.	9.5 (5.5)	2.1 (0.6)	14.6 (5.1)	2.6 (0.3)	17.1 (7.4)	2.7 (0.5)	12.8 (5.1)	2.4 (0.4)	**19.2 (3.4)*****	**2.9 (0.2)****
95th perc.	24.3 (13.0)	3.1 (0.5)	37.7 (6.6)	**3.6 (0.1)***	**44.8 (16.9)***	**3.7 (0.3)****	33.7 (7.5)	3.4 (0.2)	**42.9 (4.9)*****	**3.7 (0.1)****

Assessment of summary statistics: The established criteria for pursuing a suitable summary statistic were in decreasing order of significance: (1) variance homogeneity, (2) normality and (3) uncertainty of estimates. The first two criteria were directly derived from the assumptions underlying the linear mixed-effects model [16].

For each study and summary statistic model (1) was fitted and the standardized residuals were calculated. The variance homogeneity assumption may be violated in two distinct ways; (1a) the variance does not remain constant over the range of estimated mean values and (1b) the variance does not remain constant across dose groups. Accordingly, the assessment of the variance homogeneity assumption was 2-fold. First, to examine a possible violation as described in criterion (1a) the square root of the absolute value of the standardized residuals were fitted *via* a linear regression model using the fitted values and the p values of the slopes were calculated. Second, regarding criterion (1b) Brown–Forsythe's test (also known as the modified Levene's test) [17], which is robust to possible departures from an underlying normal distribution, was applied to the standardized residuals and the p values were calculated. The normality assumption (criterion 2) was evaluated using Shapiro–Wilk's test [18] applied to the standardized residuals and the p values were extracted. These three methods were applied to each compound and summary statistic individually and are illustrated in Fig. 2.

The extracted p values can be used as a measure of the relative performance among the different candidate statistics. As an example we consider the group variances. Even if all dose groups should have the same underlying true variance, the sampling distribution will most often be fairly right-skewed due to the inherent random variation in the data. This implies that some observed group variances will be somewhat larger than others. Applying Brown–Forsythe's test takes the inherent variation of data into account and the p values thus reflect how well the observed treatment variances conform to the distribution that is expected under the hypothesis of variance equality. High p values suggest that the model assumptions are valid while low p values may indicate that the model assumptions are violated. As we are interested in the summary statistic that most consistently meet the model assumptions, the summary statistics associated with high p values are favoured over summary statistics that are associated with low p values.

It is important to keep in mind that a p value exceeding the significance level does not guarantee that the null hypothesis is actually true [19,20]. In consequence the extracted p values should not be used to judge the significance of the hypothesis of interest. As the extracted p values are not used in the framework of hypothesis testing (*i.e.* no hypothesis tests are conducted based on these p values), there is no need to adjust for multiple testing.

Some summary statistics are more precisely estimated than others and a way to quantify this is to assess the variance of the summary statistics (criterion 3). The within-sample distributions are positively skewed and bear some resemblance with

Table 2c
Obtained % tail DNA after administration of CAS no. 115-96-8 (Tris(2-chlorethyl) phosphate) and CAS no. 598-55-0 (Methylcarbamate) in study 2. Average % tail DNA with SD in parentheses. The values were calculated by first averaging the two summary statistics for each animal and from these values the average and SD were calculated. There were five animals in each dose group and 200 cells were scored for each mouse. Data were analyzed by means of a linear mixed-effects model as defined in model (1) with Dunnett's test to compare the dose groups to their corresponding control. Values in bold indicate a significant difference. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

CAS no.	0 mg/kg bw		500 mg/kg bw		1000 mg/kg bw		1500 mg/kg bw		Positive control (EMS)	
115-96-8	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)
Mean	6.1 (2.9)	0.7 (0.6)	12.0 (5.7)	1.2 (0.6)	8.1 (4.2)	1.1 (0.7)	**12.7 (3.8)*****	**1.8 (0.5)*****	**12.2 (1.8)*****	**1.7 (0.3)****
log(mean)	1.7 (0.5)	–	2.4 (0.4)	–	2.0 (0.4)	–	**2.5 (0.3)*****	–	**2.5 (0.1)*****	–
Median	2.7 (1.7)	0.9 (0.6)	4.2 (3.1)	1.3 (0.6)	**4.2 (3.2)***	**1.2 (0.6)****	**7.3 (3.3)*****	**1.9 (0.5)*****	**7.6 (1.7)*****	**2.0 (0.2)*****
65th perc.	4.4 (2.6)	1.3 (0.6)	8.6 (6.9)	1.9 (0.7)	**6.4 (4.5)***	**1.7 (0.6)****	**10.9 (4.0)*****	**2.3 (0.3)*****	**10.8 (2.2)*****	**2.3 (0.2)*****
75th perc.	6.4 (4.0)	1.7 (0.6)	14.5 (11.2)	2.5 (0.7)	8.8 (5.5)	2.0 (0.5)	**14.9 (5.1)****	**2.6 (0.3)****	**13.9 (2.4)*****	**2.6 (0.2)****
85th perc.	9.5 (5.5)	2.1 (0.6)	23.3 (13.8)	3.0 (0.6)	14.8 (7.8)	2.6 (0.5)	**23.7 (6.5)***	**3.1 (0.3)***	**19.2 (3.4)*****	**2.9 (0.2)****
95th perc.	24.3 (13.0)	3.1 (0.5)	51.9 (23.8)	3.8 (0.5)	30.2 (12.4)	3.3 (0.4)	**45.2 (9.7)****	**3.8 (0.2)****	**42.9 (4.9)*****	**3.7 (0.1)****

CAS no.	0 mg/kg bw		500 mg/kg bw		1000 mg/kg bw		2000 mg/kg bw		Positive control (EMS)	
598-55-0	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)
Mean	6.2 (3.0)	0.7 (0.4)	**10.5 (1.9)***	**1.5 (0.2)*****	**13.4 (3.8)*****	**1.5 (0.3)*****	8.6 (2.3)	1.1 (0.2)	**12.2 (1.8)*****	**1.7 (0.3)*****
log(mean)	1.7 (0.5)	–	**2.3 (0.2)****	–	**2.5 (0.2)*****	–	2.1 (0.3)	–	**2.5 (0.1)*****	–
Median	2.4 (1.1)	0.8 (0.4)	4.4 (1.0)	**1.4 (0.2)****	**6.5 (4.0)***	**1.6 (0.3)*****	3.4 (0.7)	1.2 (0.2)	**7.6 (1.7)*****	**2.0 (0.2)*****
65th perc.	3.7 (1.7)	1.2 (0.5)	7.6 (1.8)	**2.0 (0.2)*****	**9.8 (4.1)****	**2.1 (0.3)****	5.9 (1.9)	1.7 (0.3)	**10.8 (2.2)*****	**2.4 (0.2)*****
75th perc.	5.7 (2.9)	1.6 (0.5)	**11.9 (2.6)***	**2.4 (0.2)*****	**14.0 (4.5)*****	**2.5 (0.3)*****	8.6 (3.7)	2.1 (0.4)	**13.9 (2.4)*****	**2.6 (0.2)*****
85th perc.	9.9 (5.1)	2.1 (0.5)	**19.8 (4.0)***	**2.9 (0.2)****	**24.6 (7.3)*****	**3.2 (0.3)*****	15.0 (6.3)	2.6 (0.4)	**19.2 (3.4)*****	**2.9 (0.2)****
95th perc.	26.3 (15.8)	3.0 (0.7)	44.7 (13.1)	**3.7 (0.3)***	**55.8 (12.2)****	**4.0 (0.2)****	37.6 (13.1)	3.5 (0.3)	**42.9 (4.9)***	**3.7 (0.1)***

Table 2d
Obtained % tail DNA after administration of CAS no. 636-97-5 (4-nitrobenzoic hydrazide), CAS no. 85-28-9 (4-chloro-2-hydroxy-4-methoxybenzophenone) and CAS no. 13674-87-8 (Tris[2-chloro-1-(chloromethyl)ethyl]phosphate) in study 3. Average % tail DNA with SD in parentheses. The values were calculated by first averaging the two summary statistics for each animal and from these values the average and SD were calculated. There were five animals in each dose group and 200 cells were scored for each mouse. Data were analyzed by means of a linear mixed-effects model as defined in model (1) with Dunnett's test to compare the dose groups to their corresponding control. Values in bold indicate a significant difference. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

CAS no.	0 mg/kg bw		12.5 mg/kg bw		25 mg/kg bw		50 mg/kg bw		Positive control (EMS)	
636-97-5	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)
Mean	8.5 (2.2)	0.9 (0.3)	9.3 (1.7)	0.9 (0.2)	9.7 (2.0)	1.3 (0.3)	9.5 (3.6)	0.9 (0.4)	**19.0 (2.5)*****	**2.2 (0.2)*****
log(mean)	2.1 (0.3)	–	2.2 (0.2)	–	2.2 (0.2)	–	2.2 (0.4)	–	**2.9 (0.1)*****	–
Median	2.7 (0.7)	1.0 (0.3)	2.3 (0.4)	0.8 (0.2)	2.5 (0.6)	1.3 (0.4)	2.6 (0.8)	0.9 (0.3)	**12.1 (1.9)*****	**2.5 (0.1)*****
65th perc.	4.9 (1.4)	1.5 (0.3)	4.4 (1.2)	1.4 (0.2)	6.9 (4.7)	2.0 (0.4)	4.8 (2.0)	1.5 (0.4)	**17.7 (3.3)*****	**2.8 (0.2)*****
75th perc.	7.7 (2.3)	2.0 (0.3)	7.3 (1.7)	1.9 (0.2)	10.2 (6.3)	2.5 (0.5)	8.4 (5.3)	2.7 (0.7)	**24.8 (3.8)*****	**3.2 (0.1)*****
85th perc.	14.8 (4.8)	2.6 (0.3)	18.1 (4.9)	2.8 (0.2)	19.4 (5.2)	3.1 (0.4)	19.0 (12.1)	3.8 (0.4)	**37.3 (6.5)*****	**3.6 (0.2)*****
95th perc.	40.2 (12.7)	3.6 (0.3)	44.9 (10.8)	3.8 (0.2)	44.2 (4.0)	4.0 (0.2)	46.9 (17.4)	4.0 (0.2)	**66.7 (11.4)*****	**4.2 (0.2)*****

CAS no.	0 mg/kg bw		500 mg/kg bw		1000 mg/kg bw		1500 mg/kg bw		Positive control (EMS)	
85-28-9	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)
Mean	8.4 (1.0)	1.1 (0.2)	9.8 (0.6)	1.1 (0.1)	9.1 (1.2)	1.1 (0.2)	10.0 (3.6)	1.0 (0.2)	**19.0 (2.5)*****	**2.2 (0.2)*****
log(mean)	2.1 (0.1)	–	2.3 (0.1)	–	2.2 (0.1)	–	2.2 (0.3)	–	**2.9 (0.1)*****	–
Median	4.0 (0.6)	1.4 (0.2)	3.1 (0.1)	1.1 (0.1)	3.1 (0.3)	1.1 (0.1)	2.9 (0.4)	1.1 (0.2)	**12.1 (1.9)*****	**2.5 (0.1)*****
65th perc.	6.0 (0.7)	1.8 (0.1)	5.5 (0.3)	1.7 (0.1)	5.3 (1.0)	1.6 (0.2)	6.2 (2.5)	1.7 (0.3)	**17.7 (3.3)*****	**2.8 (0.2)*****
75th perc.	8.7 (0.9)	2.1 (0.1)	9.3 (1.2)	2.2 (0.1)	8.9 (2.6)	2.1 (0.3)	10.1 (4.3)	2.2 (0.4)	**24.8 (3.8)*****	**3.2 (0.1)*****
85th perc.	14.6 (3.1)	2.6 (0.2)	18.2 (3.8)	2.9 (0.2)	15.2 (4.7)	2.7 (0.3)	19.7 (10.8)	2.8 (0.5)	**37.3 (6.5)*****	**3.6 (0.2)*****
95th perc.	31.9 (6.6)	3.4 (0.2)	47.9 (3.3)	**3.8 (0.1)***	41.3 (6.8)	3.7 (0.2)	45.7 (21.6)	3.7 (0.4)	**66.7 (11.4)*****	**4.2 (0.2)*****

CAS no.	0 mg/kg bw		225 mg/kg bw		450 mg/kg bw		900 mg/kg bw		Positive control (EMS)	
13674-87-8	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)
Mean	8.4 (1.0)	1.1 (0.2)	9.2 (2.4)	1.2 (0.4)	9.7 (1.3)	1.0 (0.2)	**14.2 (2.3)*****	**1.7 (0.3)****	**19.0 (2.5)*****	**2.2 (0.2)*****
log(mean)	2.1 (0.1)	–	2.2 (0.3)	–	2.2 (0.1)	–	**2.6 (0.2)*****	–	**2.9 (0.1)*****	–
Median	4.0 (0.6)	1.4 (0.2)	4.1 (1.4)	1.4 (0.3)	3.0 (0.4)	1.1 (0.1)	**7.2 (2.8)****	**1.9 (0.5)***	**12.1 (1.9)*****	**2.5 (0.1)*****
65th perc.	6.0 (0.7)	1.8 (0.1)	6.9 (2.9)	1.9 (0.4)	5.6 (1.0)	1.7 (0.2)	**11.9 (3.2)*****	**2.4 (0.3)*****	**17.7 (3.3)*****	**2.8 (0.2)*****
75th perc.	8.7 (0.9)	2.1 (0.1)	9.7 (4.2)	2.2 (0.4)	9.3 (1.8)	2.2 (0.2)	**17.7 (4.1)*****	**2.8 (0.2)*****	**24.8 (3.8)*****	**3.2 (0.1)*****
85th perc.	14.6 (3.1)	2.6 (0.2)	16.2 (5.0)	2.7 (0.3)	18.2 (4.8)	2.8 (0.2)	**27.1 (4.7)*****	**3.3 (0.2)*****	**37.3 (6.5)*****	**3.6 (0.2)*****
95th perc.	31.9 (6.6)	3.4 (0.2)	38.5 (11.4)	3.6 (0.3)	**44.8 (5.9)***	**3.8 (0.1)***	**58.5 (6.5)*****	**4.1 (0.1)*****	**66.7 (11.4)*****	**4.2 (0.2)*****

a log-normal distribution. Therefore, we calculated the asymptotic variance of the summary statistics of a log-normal distribution with parameters $\mu = 2$, $\sigma^2 = 2$ and $n = 100$ [21] together with the variance of the summary statistics under a logarithmic transformation as seen in Table 2. The variance of log(mean) is not provided since it does not readily compare to the variance of the other summary statistics. These results are approximate and serve to illustrate the difference in the precision of the various summary statistics.

Power calculations: Data from all 11 studies were combined (combining control groups but otherwise distinguishing different dose groups from different studies) and fitted model (1). The variance components of the variations between animals and within animal were extracted and used to calculate power of the test of dose at a 5% significance level. Rather than specifying a specific set of treatment means for all four dose groups, power was calculated for a minimum difference, D, between any two treatments [15]. This is a conservative approach assuming that the mean of the two remaining groups equal the overall average. This implies that the actual power is equal to or greater than what is obtained from the present power calculations. The power curves in Fig. 3 were calculated in terms of fold changes, where the minimum difference, D, is expressed as a ratio, *e.g.* the ratio of the response of the highest dose

Table 2e
Obtained % tail DNA after administration of CAS no. 43100-38-5 (4-*tert*-Butylbenzoic hydrazide A) and CAS no. 60965-26-6 (2-Bromo-2′,4′-dimethoxyacetophenone) in study 3. Average % tail DNA with SD in parentheses. The values were calculated by first averaging the two summary statistics for each animal and from these values the average and SD were calculated. There were five animals in each dose group and 200 cells were scored for each mouse. Data were analyzed by means of a linear mixed-effects model as defined in model (1) with Dunnett's test to compare the dose groups to their corresponding control. Values in bold indicate a significant difference. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

CAS no.	0 mg/kg bw		62.5 mg/kg bw		125 mg/kg bw		250 mg/kg bw		Positive control (EMS)	
43100-38-5	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)
Mean	8.4 (1.0)	1.1 (0.2)	9.9 (1.6)	1.1 (0.2)	**12.3 (2.8)****	1.3 (0.3)	**14.3 (1.5)*****	**1.7 (0.2)*****	**19.0 (2.5)*****	**2.2 (0.2)*****
log(mean)	2.1 (0.1)	–	2.3 (0.2)	–	**2.5 (0.2)*****	–	**2.6 (0.1)*****	–	**2.9 (0.1)*****	–
Median	4.0 (0.6)	1.4 (0.2)	2.8 (0.4)	**1.0 (0.2)***	4.0 (1.5)	1.3 (0.4)	**7.6 (0.9)*****	**2.0 (0.1)*****	**12.1 (1.9)*****	**2.5 (0.1)*****
65th perc.	6.0 (0.7)	1.8 (0.1)	5.7 (1.1)	1.7 (0.2)	8.1 (3.2)	2.0 (0.4)	**12.1 (0.5)*****	**2.5 (0.1)*****	**17.7 (3.3)*****	**2.8 (0.2)*****
75th perc.	8.7 (0.9)	2.1 (0.1)	11.0 (2.6)	2.4 (0.2)	13.8 (6.6)	2.5 (0.5)	**17.5 (2.0)*****	**2.8 (0.1)*****	**24.8 (3.8)*****	**3.2 (0.1)*****
85th perc.	14.6 (3.1)	2.6 (0.2)	20.9 (4.6)	**3.0 (0.2)***	**23.4 (8.7)***	**3.1 (0.4)***	**28.7 (2.8)*****	**3.3 (0.1)*****	**37.3 (6.5)*****	**3.6 (0.2)*****
95th perc.	31.9 (6.6)	3.4 (0.2)	44.6 (9.9)	**3.8 (0.2)***	**56.5 (13.9)*****	**4.0 (0.2)*****	**55.5 (8.7)*****	**4.0 (0.2)*****	**66.7 (11.4)*****	**4.2 (0.2)*****

CAS no.	0 mg/kg bw		45 mg/kg bw		90 mg/kg bw		180 mg/kg bw		Positive control (EMS)	
60965-26-6	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)	Raw data	log(data)
Mean	8.4 (1.0)	1.1 (0.2)	9.7 (0.9)	1.2 (0.1)	**14.3 (7.5)***	**1.8 (0.5)****	12.9 (0.8)	**1.8 (0.1)*****	**19.0 (2.5)*****	**2.2 (0.2)*****
log(mean)	2.1 (0.1)	–	2.3 (0.1)	–	**2.6 (0.5)***	–	**2.5 (0.1)***	–	**2.9 (0.1)*****	–
Median	4.0 (0.6)	1.4 (0.2)	3.4 (0.4)	1.2 (0.1)	8.9 (6.9)	**2.0 (0.6)***	7.3 (1.2)	**2.9 (0.2)***	**12.1 (1.9)*****	**2.5 (0.1)*****
65th perc.	6.0 (0.7)	1.8 (0.1)	5.9 (1.0)	1.7 (0.2)	13.2 (10.2)	**2.4 (0.6)***	11.8 (0.9)	**2.5 (0.1)****	**17.7 (3.3)*****	**2.8 (0.2)*****
75th perc.	8.7 (0.9)	2.1 (0.1)	9.0 (1.3)	2.2 (0.2)	19.0 (14.5)	**2.7 (0.6)***	15.1 (1.1)	**2.7 (0.1)***	**24.8 (3.8)*****	**3.2 (0.1)*****
85th perc.	14.6 (3.1)	2.6 (0.2)	16.8 (2.8)	2.8 (0.2)	27.3 (16.7)	**3.2 (0.6)***	24.1 (3.4)	**3.2 (0.1)***	**37.3 (6.5)*****	**3.6 (0.2)*****
95th perc.	31.9 (6.6)	3.4 (0.2)	46.8 (6.6)	3.8 (0.1)	**50.7 (17.5)***	**3.9 (0.4)***	44.9 (10.8)	3.8 (0.2)	**66.7 (11.4)*****	**4.2 (0.2)*****

group to the control group. Power curves were generated for 1.5-, 2- and 2.5-fold changes for the median of the log-transformed data when 50 and 100 cells were scored per gel, respectively.

All statistical analyses were performed using R, version 3.0.2 [14].

3. Results

3.1. Histopathological examination and the incidence of highly damaged cells

No treatment-related effects were observed in the histological examination of testes. Occasional multinucleated germ cells were seen in one testis of a mouse exposed to the highest dose of 4-*tert*-butylbenzoic hydrazide. These appeared to be fused groups of elongated spermatids and are known as a sign of degeneration [22]. This specific morphology is a rare finding and was not seen in any controls or other exposure groups, but it cannot be determined whether this was dose-related or not. Testicular toxicity of the structurally related 4-*tert*-butylbenzoic acid has been described previously [23,24], but at higher doses than applied here. Prolonged exposure to 4-*tert*-butylbenzoic hydrazide may reveal clearer signs of testicular toxicity than the short exposure period applied in this study.

The minimum of highly damaged cells (>80% tail DNA) per gel was 0% and the maximum was 8%. Per animal the minimum was 0% and maximum was 6% (positive control animal and highest dosed animal of CAS no. 43100-38-5). The values for the control dose groups (5 animals) varied between 0.5 and 1.5%, positive control animals between 0.6 and 2.4% and for the dosed groups of all the CAS nos. between 0 and 2.7%. No treatment related increases of highly damaged cells were observed.

3.2. Levels of DNA damage

Table 2 show the % tail DNA for selected summary statistics extracted for each study. In most of the 11 studies significance was present for some summary statistics while not for others implying that the summary statistics may capture the effect induced by the treatments with varying efficiency. In a few cases the log-transformation stabilized the results in the sense that significance was obtained for a broader range of summary statistics than for the untransformed data rendering the choice of a specific summary statistic less important.

3.3. Assessment of summary statistics

Suitable statistics to summarize the within-sample distributions were evaluated according to the following criteria:

Variance homogeneity (constant variance over the range of estimated mean values): Fig. 2a shows the *p* values obtained from the least squares fit of the square root of the absolute value of the standardized residuals *versus* the fitted values. The null hypothesis is that the slope of the regression line is 0, which occurs in case of constant variance. High *p* values for testing the slope thus support that the assumption of constant variance are valid whereas low *p* values may indicate a violation. The grey line outlines the median for each summary statistic.

The summary statistics of the raw data were associated with *p* values that were increasing with increasing percentiles. The mean of the raw data was associated with a median *p* value that was higher than what was observed for most of the percentiles. In general, the summary statistics of the log-transformed data resulted in higher *p* values, which was seen especially for the mean and the median.

Variance homogeneity (constant variance across dose groups): Fig. 2b shows the *p* values from the Brown–Forsythe's test. The null hypothesis is that the variance of the standardized residuals are constant across different dose groups, and again high *p* values support the validity of this assumption whereas low *p* values may indicate a violation.

Overall, no difference appeared between the summary statistics obtained from raw and log-transformed data, respectively. The percentiles in the range around the 75th percentile of both raw and log-transformed data as well as the median of the raw data were associated with lower *p* values.

Normality: The *p* values from the Shapiro–Wilk's test are given in Fig. 2c. The null hypothesis is that the standardized residuals are normally distributed.

The mean and the 95th percentile were associated with high *p* values whereas fairly low *p* values were seen for the remaining summary statistics derived from the raw data. In general, the

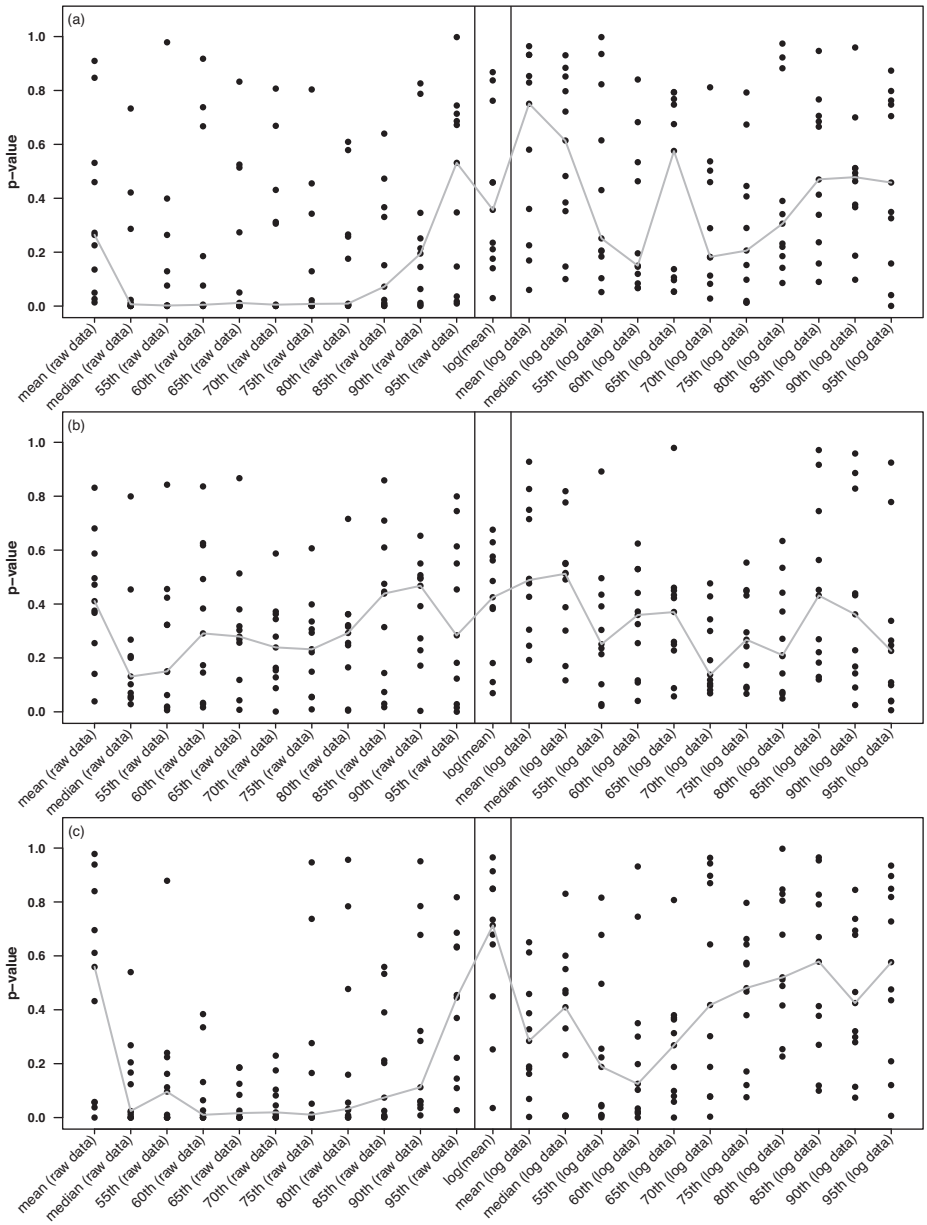

Fig. 2. Assessment of the variance homogeneity and the normality assumption. The depicted p values are concerned with (a) variance homogeneity over the range of estimated mean values, (b) variance homogeneity across dose groups and (c) normality. The median p values are given by the grey lines and the two vertical lines separate the summary statistics of the raw data, the log(mean) and the summary statistics of the log-transformed data. The 'xth (raw data)' and 'xth (log data)' are short for the xth percentile of the raw data and of the log-transformed data, respectively.

p values associated with the summary statistics of the log-transformed data had lower levels but were more consistent than the summary statistics of the raw data. The log(mean) was also associated with a relatively high p value.

Uncertainty of estimates: The asymptotic variances of the summary statistics are given in Table 3. The first column outlines the variances assuming that the raw data are log-normally distributed. The second column contains the variances in case of a logarithmic

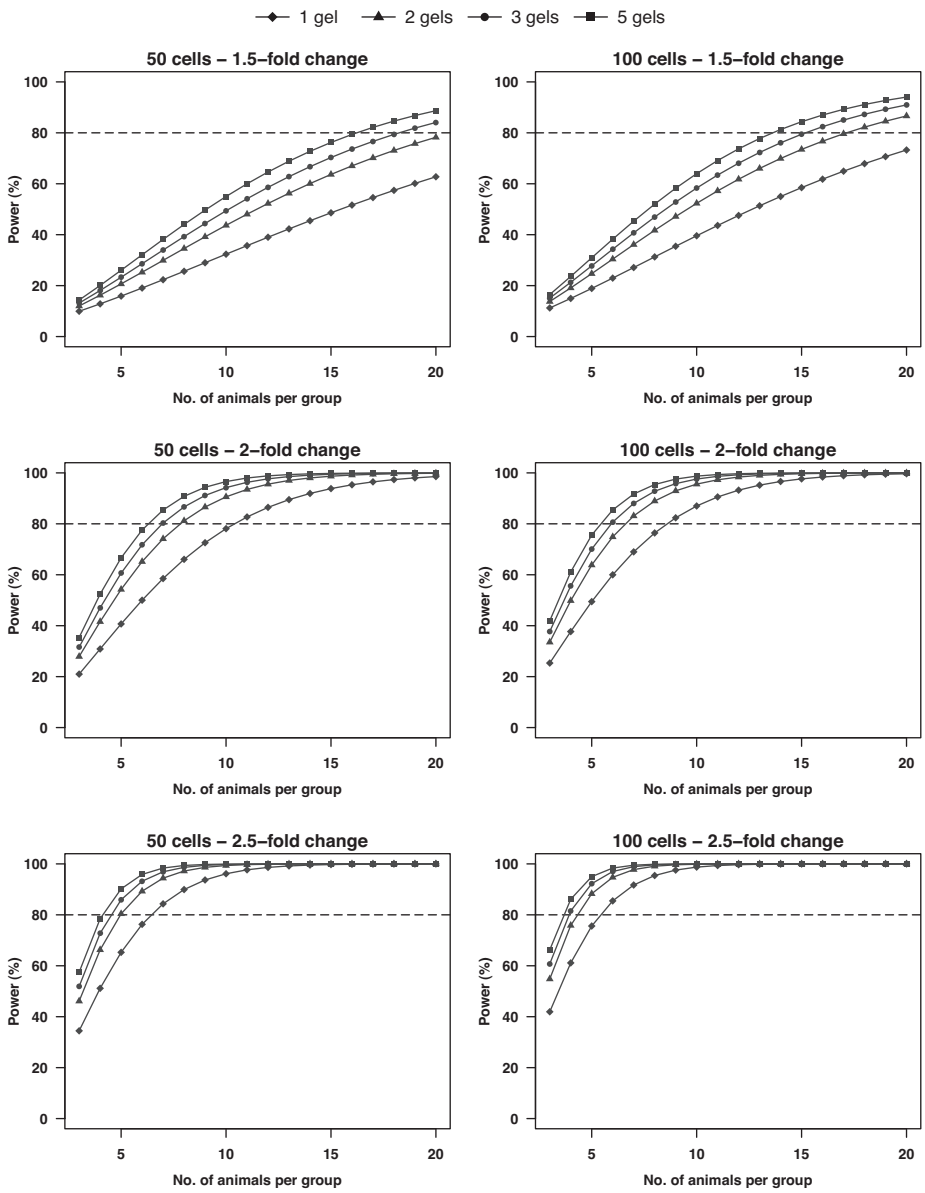

Fig. 3. Power curves outlining the number of animals per group and gels per animal required to detect certain fold changes with a power of 80% (dotted line) when 50 cells (first column) and 100 cells (second column) are scored per gel, respectively. The power calculations apply when the summary statistic is the median of the log-transformed data.

transformation. The variances are comparable within each column and should not be compared across the columns.

Assuming a log-normal distribution (resembling the raw data) the median has the lowest variance. The variance increases vastly with increasing percentiles. The variance of the 60th percentile is approximately twice the size of the variance of the median and the 90th and the 95th percentile is around 70 and 300 times the variance of the median. The variance of the mean is comparable to the variance of the 80th percentile, which are noticeably greater than the variance of the median.

Under a log-transformation, the mean is associated with the smallest variance whereas the variance of the summary statistics in the range of the median to the 75th percentile is within twice

M.K. Hansen et al. / Mutation Research 774 (2014) 29–40 37

Table 3
The variance of each summary statistic under the assumption that the within-sample distributions for each gel are log-normally distributed with $\mu = 2$, $\sigma^2 = 2$ and $n = 100$. The variances are multiplied by a factor of 50 for readability.

Summary statistic	Variance assuming a log-normal distribution (resembling raw data[a])	Variance assuming a normal distribution (resembling log(data)[b])
Mean	64.2	1.0
Median	4.3	1.6
55th perc.	6.1	1.6
60th perc.	9.0	1.6
65th perc.	13.4	1.7
70th perc.	20.8	1.7
75th perc.	34.0	1.9
80th perc.	60.0	2.0
85th perc.	119.6	2.3
90th perc.	298.0	2.9
95th perc.	1272.5	4.7

[a] The calculation of the variance is based on the assumption that the raw data follow a log-normal distribution.
[b] Log-normally distributed data subjected to a log-transformation are normally distributed.

the size of the variance of the mean. For the higher percentiles the variance is more steeply increasing and the variance of the 90th and 95th percentiles is approximately 3 and 4.5 times the variance of the mean, respectively.

3.4. Power calculations

The evaluation of the proposed summary statistics is given in the Section 5 and points to the median of the log-transformed data as the most expedient. Therefore, power curves are provided for this measure.

The power calculations were based on parameter estimates extracted from a fit of model (1) to the combined data from all 11 studies when 50 and 100 cells were scored, respectively. When 50 cells were scored the overall average of the response of the control groups were $\mu = 0.98$ and the estimated variance components were

$$\hat{\sigma}_A^2 = 0.09, \quad \hat{\sigma}^2 = 0.11.$$

When 100 cells were scored then $\mu = 0.96$ and

$$\hat{\sigma}_A^2 = 0.08, \quad \hat{\sigma}^2 = 0.09.$$

Since the analysis were performed on the summary statistics instead of the raw within-sample observations, $\hat{\sigma}^2$ is a pooled estimate containing both a contribution from gel-to-gel variation and the within-sample variation. The size of the estimated variance component reflecting the animal-to-animal variation, $\hat{\sigma}_A^2$ was of the same magnitude as the size of the estimated error variance, $\hat{\sigma}^2$, implying that the animal-to-animal variation indeed should be accounted for in the statistical analysis as is the case with the mixed-effects model .

Fig. 3 shows the power for a 1.5-, 2- and 2.5-fold change of one of the dose groups compared to the control group when 50 and 100 cells were scored. A power of 80% is outlined by a dotted line. In all cases the use of 2 gels per animal compared to 1 gel per animal improved the power noticeably, whereas the impact of a further increase in the employed number of gels was somewhat diminished. In general, the increase in power using 2 gels instead of 1 gel was of the same magnitude as using 5 gels instead of 2 gels.

Small differences, *e.g.* a 1.5-fold change, were associated with low power (< 80%) for numbers of animal less than 10 irrespective of the number of gels used, *e.g.* to detect a 1.5-fold change with a power of 80% using 5 gels per animal and 100 cells per gel required 14 animals per group. Power was increasing with increasing fold changes and for a 2-fold change 5 animals were sufficient to achieve a power of 80% using 2 gels scoring 100 cells per gel. To detect a 2.5-fold change with a power of 80% only 3 animals and 2 gels per animal were needed when 100 cells per gel were scored. Power was in general increasing with the number of cells that were scored per gel. Increasing the number of cells from 50 to 100 cells per gel increased power.

4. Interpreting the results when data are log-transformed

One natural way to formulate the outcome of a linear mixed-effects model is to express it in terms of the difference between the response of the different dose groups and the response of the control group, *e.g.* as

$$\mu_{\text{highest dose group}} - \mu_{\text{control group}} \tag{2}$$

These differences are invariant to the level of the control group. Now assume that $X \sim N(\mu, \sigma^2)$ and $Y = \exp(X)$, so that $Y \sim LN(\mu, \sigma^2)$, *i.e.* Y follows a log-normal distribution. In this case the differences defined in Eq. (2) for the normally distributed data, X, translates into a ratio for the log-normally distributed data, Y, namely

$$\mu_j - \mu_i = ln\,\frac{E(Y)_j}{E(Y)_i} \tag{3}$$

where μ_i denotes the mean of the normally distributed data, X, for the ith treatment group and $E(Y)_i$ denotes the expected value (the mean) of the log-normally distributed data, Y, for the ith treatment group. The most suitable summary statistic was in this study identified as the median of the log-transformed data. The logarithmic transformation is monotonic, that is, the order of the data is preserved after transformation, and therefore

$$\log(\text{median}(\text{data})) = \text{median}(\log(\text{data})) \tag{4}$$

Based on the earlier shown findings and further graphical investigations (not shown) it is assumed in the following that $\text{median}(\text{data}) \sim LN(\mu, \sigma^2)$ (corresponding to Y) and that $\text{median}(\log(\text{data})) \sim N(\mu, \sigma^2)$ (corresponding to X). This means that differences obtained from the statistical analysis of the log-transformed data translates into x-fold changes for the raw data. Consider the following example: from the analysis of the log-transformed data a difference of $\mu_{\text{highest dose group}} - \mu_{\text{control group}} = 0.69$ is obtained. From Eq. (3) we see that $E(Y)_{\text{highest dose group}}/E(Y)_{\text{control group}} = \exp(0.69) = 2$, meaning that the obtained result corresponds to a 2-fold change relative to the control group on the original scale of the raw data. An important feature is that the obtained fold changes are irrespective of the actual response level of the different groups.

These properties also apply to the power curves shown in Fig. 3. This implies that power calculations can be made for certain fold

changes without making any assumptions of the actual levels of the different treatment groups, which is exactly what is reflected in the shown power curves. It would also be possible to specify fold changes for normally distributed data, but in this case the power will change with the response level of the baseline group (*e.g.* the control group) and assumptions about this level would have to be made in order to calculate power.

It may be desirable to report the absolute group levels instead of or as a supplement to the fold changes. A simple method is to report the back-transformed group means

$$\exp(\mu_i)$$

and the appertaining confidence interval is obtained by back-transforming the limits of the confidence interval of the group mean [25]. However, it is important to be aware that this back-transformation does not yield the familiar arithmetic mean which is normally used but rather provides the geometric mean [25]. The geometric mean is not easily interpreted but for log-normally distributed data it is a good estimate of the median. The median may be more sensible to report than the mean in case of skewed distributions as is the case of log-normally distributed data.

There may be some practical issues extracting a summary statistic of the log-transformed data since certain summary statistics of the raw data such as the mean and median may be automatically provided by the Comet assay software. However, in accordance with Eq. (4) the median of the log-transformed data can be calculated by taken the logarithm of the median of the raw data. The expression is in general valid for all percentiles but does not apply to the mean.

5. Discussion

One aim of this study was to identify a statistic adequately summarizing the empirical distribution of the % tail DNA of testicular samples of mice obtained for each gel in a Comet assay study. The criteria used in this appraisal were considering the assumptions made by the appropriate statistical analysis together with the robustness towards the estimation of the summary statistics in question. A second aim was to provide power curves outlining the appropriate number of animals and gels to use when analysing testicular samples in the Comet assay.

For both aims the results are highly influenced by the modelling approach. A natural consequence of the setup of the Comet assay is that the obtained data are structured hierarchically, which are to be modelled appropriately. Also, animal should be modelled as a random effect to ensure proper inference, *i.e.* that the conclusions drawn can be generalized. The inherent nature of the Comet assay setup therefore suggests the use of a linear mixed-effects model as defined in model (1).

Different proposals for a suitable summary statistic have been proposed, some of which are considering power [6] and higher sensitivity [7] based on actual data. Others discuss the presumptive behaviour of the summary statistics regarding normality [2] and robustness of estimates [4] grounded on theoretical considerations. The current study bridges these approaches using actual data for examining the prerequisites underlying valid inference from the statistical analysis including the issues such as normality and constant variance, and also addresses robust estimation of the summary statistics. The diversity in objectives partly explains the different recommendations outlined. Spurious conclusions may be drawn when the requirements of the statistical analysis in question are not met. Our view is therefore that it is more crucial to examine if the assumptional constraints are fulfilled and that considerations regarding desirable properties such as sensitivity and power is only of secondary importance.

The assessment of variance homogeneity and normality revealed that the summary statistics met the standards diversely. The summary statistics most consistently complying with these criteria were the mean, 90th and 95th percentile of the raw data, the log(mean) and the mean, median, 85th and 90th percentile of the log-transformed data. The variance homogeneity (criterion 1) was considered the most important in this study followed by normality (criterion 2) and the following issues only apply to these statistics since they most optimally complied with these two criteria. Although the variances listed in Table 2 are specific to the log-normal distribution, they reflect the diversity obtained for the various summary statistics when the underlying distribution is highly skewed. It is evident from Table 3 that the variance increases immensely with increasing percentiles. The log-transformation of data stabilizes the variance, still higher percentiles are associated with moderately higher variances compared to the mean or median. Moreover, as pointed out by Duez et al. [7], higher percentiles may more rapidly saturate than the mean or lower percentiles. This is a likely consequence of the nature of the percentage data since the % tail DNA by definition cannot exceed 100%. Therefore, it is possible that the lower and middle part reshapes more distinctly than the upper part in response to administration of a genotoxic compound.

Another consideration is that when data are log-transformed the natural interpretation of the results changes as described in the previous section. The interpretation in terms of fold changes relative to the control group is applicable when the summary statistic is any of the percentiles (including the median). This does not apply when the summary statistic is the mean of the log-transformed data or the log(mean) in which case it is not clear how to interpret the estimates from the statistical analysis. Altogether, from the current data and the priority of the criteria mentioned we conclude that the median of the log-transformed data was the most suitable summary statistic since it most consistently conformed to the range of issues considered. Alternatively, the mean of the raw data met the criteria almost equally well. Although the interpretation of the estimates from the statistical analysis differ according to which of the two summary statistics that are used, the interpretation is well-defined in both cases. Interestingly, the median of the raw data did not meet the assumptions of the statistical model.

Power analyses performed by Smith et al. [8] reveal that noticeable differences in power are obtained for different types of tissues from rats. Power calculations of cells from the tail vein blood give considerably lower power compared to cells from the stomach, liver and bone marrow. The number of animals required to obtain 80% power for testicular cells in the present study is comparable to the number of animals required for stomach, liver and bone marrow cells for a 2-, 2.5- and 3-fold change as reported by Smith et al. [8]. In the present study, the use of 2 gels per animal compared to 1 gel per animal improved the power noticeably, whereas the impact of a further increase in the number of gels was somewhat diminished.

To our knowledge, only one of the 11 chemicals tested in the present study, namely acrylamide, has been tested in testicular cells in the comet assay, and limited information is available of comet assay data in other tissues and other genotoxicological tests (Table 1). In the present study, acrylamide induced an effect in testicular cells which is in line with another study where comet assay results revealed effects in gonadal sperm cells and testicular somatic cells in mice and rats [26,27]. This chemical has also induced *in vivo* genotoxic effects in other assays than the comet assay [28]. In the present study, perfluorooctane sulfonate did not induce an effect in testicular cells in mice at 300 mg/kg bw, however it has induced effects in rat bone marrow cells in the comet assay and in the micronucleus assay in female rats at 0.6 mg/kg bw/day [29]. Tris(2-chlorethyl) phosphate induced an effect in

testicular cells at 1500 mg/kg bw in the present study. This chemical was tested positive in Chinese Hamster bone marrow cells in the micronucleus assay at 400 mg/kg bw, however only two males and two females were used in the dose groups [30]. In another study in the micronucleus assay in mice, no effects were observed [31]. Methyl phosphate phosphoric acid, trimethyl ester induced an effect at 500 mg/kg bw in the present study. It was tested positive in the dominant lethal mutation assay in mice at 1000 mg/kg bw [32], but was negative in bone marrow cells in the micronucleus assay in mice [33]. Methyl carbamate induced an effect in the present study at 1000 mg/kg bw and was tested negative in induction of sex linked mutations in Drosophila [34]. In the present study, Tris[2-chloro-1-(chloromethyl)ethyl]phosphate induced an effect at 900 mg/kg bw. It did not induce chromosomal aberrations in rats after oral exposure at up to 165 mg/kg bw [35] and no clastogenic effects in mice after intraperitoneal exposure up to 350 mg/kg bw was observed [35].

In conclusion, we summarize the results from the current study regarding the design and analysis of the % tail DNA of cells from mice testicular cells in the Comet assay study. Prior to conducting the experiment a desired fold change must be specified and the number of animals to use can be identified from Fig. 3. In general, more than 10 animals per group were needed for fold changes less than 2 whereas less than 10 animals per group were sufficient for fold changes of 2 or more. Fig. 3 shows that the largest increase in power was gained when using 2 gels per animal instead of 1 gel whereas less was gained using 3 gels instead of 2 gels *etc.* The obtained within-sample distributions were most suitably summarized by the median of data subjected to the natural logarithm or alternatively the mean of the raw data. If the median of the log-transformed data is used as the summary statistic the natural logarithm of the median of the raw data yields an equivalent result. The calculated summary statistics should subsequently be used in the statistical analysis. The analysis performed should be a linear mixed-effects model with animal nested within dose to avoid inflation of type I and type II errors. The interpretation of the outcome of the statistical analysis should reflect the choice of summary statistic, *i.e.* if the median of the log-transformed data is used the estimates obtained from the statistical analysis are interpreted as described in the previous section.

Conflict of interests

There are no conflict of interests.

Acknowledgements

This study was partly funded by the Danish EPA. Vivian Jørgensen, Alicja Mortensen, Sarah Grundt Simonsen, Annette Landin, personnel in the animal facility from DTU-National Food Institute-Denmark, Charles Homsy and Françoise Soussaline, IMSTAR S.A., France, are gratefully acknowledged.

Appendix A. R code for fitting linear mixed-effects model with a nested structure in R

The following R code can be used to fit model (1) to Comet assay data. The data should be structured column-wise with a treatment column, an animal column and the response (the calculated summary statistic). The following example shows a data set with a control group and three dose groups, with 5 animals in each group and two gels per animal. A csv-file containing this data may look as

```
Dose, Animal,
Response
Control, 1, xxx
Control, 1, xxx
Control, 2, xxx
Control, 2, xxx
Control, 3, xxx
...
Dose1, 6, xxx
Dose1, 6, xxx
Dose1, 7, xxx
Dose1, 7, xxx
...
Dose3, 20, xxx
```

where `xxx` is the value of the calculated summary statistic and "," is the delimiter. Note the consecutive numbering of the animals (numbered 1–20 instead of 1–5 which is repeated for each dose group). Although it is not strictly necessary to number the animals in this way in R it is good idea since it indicates the nesting structure which otherwise inadvertently may be disregarded.

The statistical software R is available at http://www.r-project.org/. To fit model (1) in R the package `nlme` [36] or `lme4` [37] needs to be installed in addition to base R. In the following example we will use the `nlme` package. Also, we need `multcomp` [38] for performing Dunnett's test. They are installed and loaded with

```
install.packages("nlme")
library(nlme)
install.packages("multcomp")
library(multcomp)
```

The csv-file can be imported into R with

```
dat <- read.csv("nameoffile.csv", header=TRUE, sep=",", dec=".")
```

Note that if the csv-file uses an alternative separator or decimal point the relevant symbol can specified with the `sep` and `dec` arguments. To ensure that Animal is a factor

```
dat$Animal <- factor(dat$Animal)
```

Model (1) can now be fitted with

```
m1 <- lme(Response ~ Dose, random = ~ 1 | Animal, data = dat)
```

An omnibus test of Dose is performed with

```
anova(m1)
```

The estimates can be extracted with

```
fixef(m1)
```

and Dunnett's test can be performed with

```
summary(glht(m1, linfct = mcp(Dose = "Dunnett")))
```

The variance components $\hat{\sigma}_A^2$ and $\hat{\sigma}^2$ are extracted with

```
VarCorr(m1)
```

References

[1] A.M. Lynch, J.C. Sasaki, R. Elespuru, D. Jacobson-Kram, V. Thybaud, M. De Boeck, M.J. Aardema, J. Aubrecht, R.D. Benz, S.D. Dertinger, G.R. Douglas, P.A. White, P.A. Escobar, A. Fornace, M. Honma, R.T. Naven, J.F. Rusling, New and emerging technologies for genetic toxicity testing, Environ. Mol. Mutagen. 52 (2011) 205–223.
[2] D.P. Lovell, T. Omori, Statistical issues in the use of the comet assay, Mutagenesis 23 (2008) 171–182.
[3] P.E. Verde, L.A. Geracitano, L.L. Amado, C.E. Rosa, A. Bianchini, J.M. Monserrat, Application of public-domain statistical analysis software for evaluation and comparison of comet assay data, Mutat. Res.-Gen. Tox. Environ. 604 (2006) 71–82.
[4] J. Bright, M. Aylott, S. Bate, H. Geys, P. Jarvis, J. Saul, R. Vonk, Recommendations on the statistical analysis of the comet assay, Pharm. Stat. 10 (2011) 485–493.
[5] D.P. Lovell, G. Thomas, R. Dubow, Issues related to the experimental design and subsequent statistical analysis of in vivo and in vitro comet studies, Teratogen. Carcin. Mut. 19 (1999) 109–119.

[6] S.J. Wiklund, E. Agurell, Aspects of design and statistical analysis in the comet assay, Mutagenesis 18 (2003) 167–175.
[7] P. Duez, G. Dehon, A. Kumps, J. Dubois, Statistics of the comet assay: a key to discriminate between genotoxic effects, Mutagenesis 18 (2003) 159–166.
[8] C.C. Smith, D.J. Adkins, E.A. Martin, M.R. O'Donovan, Recommendations for design of the rat comet assay, Mutagenesis 23 (2008) 233–240.
[9] A. Baumgartner, E. Cemeli, D. Anderson, The comet assay in male reproductive toxicology, Cell Biol. Toxicol. 25 (2009) 81–98.
[10] G. Speit, M. Vasquez, A. Hartmann, The comet assay as an indicator test for germ cell genotoxicity, Mutat. Res.-Rev. Mutat. Res. 681 (2009) 3–12.
[11] R.R. Tice, E. Agurell, D. Anderson, B. Burlinson, A. Hartmann, H. Kobayashi, Y. Miyamae, E. Rojas, J.C. Ryu, Y.F. Sasaki, Single cell gel/comet assay: guidelines for in vitro and in vivo genetic toxicology testing, Environ. Mol. Mutagen. 35 (2000) 206–221.
[12] A. Hartmann, E. Agurell, C. Beevers, S. Brendler-Schwaab, B. Burlinson, P. Clay, A. Collins, A. Smith, G. Speit, V. Thybaud, R.R. Tice, Recommendations for conducting the in vivo alkaline Comet assay, Mutagenesis 18 (2003) 45–51.
[13] A.K. Sharma, F. Soussaline, J. Sallette, M. Dybdahl, The influence of the number of cells scored on the sensitivity in the comet assay, Mutat. Res.-Gen. Tox. Environ. 749 (2012) 70–75.
[14] R Core Team, R: A Language and Environment for Statistical Computing, Austria, Vienna, 2013.
[15] D.C. Montgomery, Design and Analysis of Experiments, 6th ed., John Wiley & Sons, Inc., USA, 2005.
[16] J.C. Pinheiro, D.M. Bates, Mixed-Effects Models in S and S-PLUS, Springer-Verlag, New York, NY, 2000.
[17] M.B. Brown, A.B. Forsythe, Robust tests for the equality of variances, J. Am. Stat. Assoc. 60 (1974) 364–367.
[18] S.S. Shapiro, M.B. Wilk, An analysis of variance test for normality (complete samples), Biometrika 52 (1965) 591–611.
[19] D.G. Altman, J.M. Bland, Absence of evidence is not evidence of absence, Br. Med. J. 311 (1995) 485.
[20] D.J. Murdoch, Y.-L. Tsai, J. Adcock, P-values are random variables, Am. Stat. 62 (2008) 242–245.
[21] D.R. Cox, D.V. Hinkley, Theoretical Statistics, Chapman and Hall, London, 1974.
[22] C.G. Print, K.L. Loveland, Germ cell suicide: new insights into apoptosis during spermatogenesis, Bioessays 22 (2000) 423–430.
[23] G.L. Hunter, P.L. Chambers, D.E. Stevenson, Studies on the oral toxicity of p-tert-butyl benzoic acid in rats, Food. Cosmet. Toxicol. 3 (1965) 289–298.
[24] M.E. Bernstein, Agents affecting the male reproductive system: effects of structure on activity, Drug Metab. Rev. 15 (1984) 941–996.
[25] M. Bland, An Introduction to Medical Statistics, 3rd ed., Oxford University Press, Oxford, 2000.
[26] L. Recio, C. Hobbs, W. Caspary, K.L. Witt, Dose-response assessment of four genotoxic chemicals in a combined mouse and rat micronucleus (MN) and Comet assay protocol, J. Toxicol. Sci. 35 (2010) 149–162.
[27] K.L. Witt, E. Livanos, G.E. Kissling, D.K. Torous, W. Caspary, R.R. Tice, L. Recio, Comparison of flow cytometry- and microscopy-based methods for measuring micronucleated reticulocyte frequencies in rodents treated with nongenotoxic and genotoxic chemicals, Mutat. Res.-Gen. Tox. Environ. 649 (2008) 101–113.
[28] A. Besaratinia, G.P. Pfeifer, DNA adduction and mutagenic properties of acrylamide, Mutat. Res.-Gen. Tox. Environ. 580 (2005) 31–40.
[29] A. Celik, D. Eke, S.Y. Ekinci, S. Yildirim, The protective role of curcumin on perfluorooctane sulfonate-induced genotoxicity: single cell gel electrophoresis and micronucleus test, Food Chem. Toxicol. 53 (2013) 249–255.
[30] M. Sala, Z.G. Gu, G. Moens, I. Chouroulinkov, In vivo and in vitro biological effects of the flame retardants tris(2,3-dibromopropyl) phosphate and tris(2-chloroethyl)orthophosphate, Eur. J. Cancer Clin. Oncol. 18 (1982) 1337–1344.
[31] M. Beth-Hübner, Toxicological evaluation and classification of the genotoxic, carcinogenic, reprotoxic and sensitising potential of tris(2-chloroethyl) phosphate, Int. Arch. Occup. Environ. Health 72 (1999) M17–M23.
[32] H. Tezuka, Y.F. Sasaki, M. Inoue, A. Uchida, M. Moriya, Y. Shirasu, Heritable translocation study in male mice with trimethyl phosphate, Mutat. Res. 157 (1985) 205–213.
[33] Z. Ni, S. Li, Y. Liu, Y. Tang, D. Pang, Induction of micronucleus by organophosphorus pesticides both in vivo and in vitro, Hua Xi Yi Ke Da Xue Xue Bao (J. West China Univ. Med. Sci.) 24 (1993) 82–86.
[34] National Toxicology Program (NTP), Toxicology and Carcinogenesis Studies of Methyl Carbamate (CAS No. 598-55-0) in F344/N Rats and B6C3F1 Mice (Gavage Studies), NC, 1987.
[35] World Health Organization (WHO), Flame Retardants: Tris(chloropropyl) phosphate and Tris(2-chloroethyl) phosphate, WHO, Geneva, Switzerland, 1998.
[36] J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, R Development Core Team, nlme: Linear and Nonlinear Mixed Effects Models, 2013.
[37] D. Bates, M. Maechler, B. Bolker, S. Walker, lme4: Linear Mixed-effects Models using Eigen and S4, 2013.
[38] T. Hothorn, F. Bretz, P. Westfall, Simultaneous inference in general parametric models, Biometr. J. 50 (2008) 346–363.

Appendix E

A national drug related problems database - evaluation of use in practice, reliability and reproducibility

---

Kjeldsen, L. J., Birkholm, T., Fischer, H., Graabæk, T., Hansen, M. K., Kibsdal, K. P., Ravn-Nielsen, L. and Truelshøj, T. H. (2014). A national drug related problems database - evaluation of use in practice, reliability and reproducibility. *International Journal of Clinical Pharmacy*, *36*, 742-749. DOI 10.1007/s11096-014-9957-2.

Appendix F

binomTools: Performing diagnostics on binomial regression models

---

Christensen, R. H. B. and Hansen, M. K. (2012). **binomTools:** Performing diagnostics on binomial regression models. R package version 1.0-1. http://CRAN.R-project.org/package=binomTools/.

Package 'binomTools'

August 29, 2013

Type Package

Title Performing diagnostics on binomial regression models

Version 1.0-1

Date 2011-08-03

Author Rune Haubo B Christensen and Merete K Hansen

Maintainer Merete K Hansen <mkh@imm.dtu.dk>

Description This package provides a range of diagnostic methods for binomial regression models.

License GPL (>= 3)

LazyLoad yes

Repository CRAN

Date/Publication 2011-08-09 11:36:34

NeedsCompilation no

R topics documented:

beetles *Mortality of confused flour beetles*

Description

Data from a study examining the response of confused flour beetles to increasing concentrations of gaseous carbon disulphide. After exposure for five hours the exact concentration of carbon disulphide was determined and the number of dead flour beetles were recorded. For each concentration of carbon disulphide duplicate batches of beetles were used.

Usage

```
data(beetles)
```

Format

A data frame with 16 observations on the following 4 variables.

`conc` Concentration of carbon disulphide (mg/l)

`rep` Replicate number

`y` Number of deaths in each dose gruop

`n` Total number of beetles in each dose group

Source

Strand, A.L. (1930) Measuring the toxicity of insect fumigants. *Industrial and Engineering Chemistry: analytical edition*, **83**, 426-431.

References

Collett, D. (2003) *Modelling binary data*. Second edition. Chapman & Hall/CRC.

empLogit *Calculates the empirical logit transform*

Description

The empirical logit transform allows for a tolerence such that infinity is not returned when the argument is zero or one.

Usage

```
empLogit(x, eps = 1e-3)
```

Arguments

x	numerical vector for which the empirical logit transform is desired
eps	numerical scalar; a tolerence to prevent infinite values

Value

the empirical logit transform of x

Author(s)

Rune Haubo B Christensen

Examples

```
## The function is currently defined as
## function (x, eps = 1e-3) log((eps + x)/(1 - x + eps))

## Lifted from example(predict.glm):
ldose <- rep(0:5, 2)
numdead <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
sex <- factor(rep(c("M", "F"), c(6, 6)))
SF <- cbind(numdead, numalive=20-numdead)
## budworm.lg <- glm(SF ~ sex*ldose, family=binomial)
## summary(budworm.lg)

empLogit(numdead/20)

## Possible usage:
## Explorative interaction plot:
interaction.plot(ldose, sex, empLogit(numdead/20))
```

---

exact.deletion *Exact deletion residuals*

Description

Function to derive exact values of deletion (leave-one-out) residuals for binomial regression models

Usage

```
exact.deletion(object)
```

Arguments

object	An object of class glm with a binomial family

Details

The ith deletion residual is calculated subtracting the deviances when fitting a linear logistic model to the full set of n observations and fitting the same model to a set of $n-1$ observations excluding the ith observation, for $i = 1, ..., n$. This gives rise to $n+1$ fitting processes and may be computationally heavy for large data sets.

Approximations to the deletion residuals, as described in Williams (1987), are provided by `rstudent`.

Inconsistency regarding the terminology implies that the deletion residuals are called different names in the litterature, including likelihood residuals, studentized residuals, externally studentized residuals, deleted studentized residuals and jack-knife residuals. Conversely, some of these terms refer to different types of residuals

Value

A vector with exact deletion residuals

Author(s)

Merete K Hansen

References

Collett, D. (2003) *Modelling binary data*. Second edition. Chapman & Hall/CRC.

Fox, J. (2002) *An R and S-Plus Companion to Applied Regression*. Sage Publ.

Williams, D. A. (1987) Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics* **36**, 181-191.

See Also

`Residuals`, `rstudent`

Examples

```
data(beetles)
beetles.glm <- glm(cbind(y, n-y) ~ log(conc), family=binomial, data=beetles)
exact.deletion(beetles.glm)
```

---

group *Group observations in a binomial glm*

Description

This function groups the observations in a binomial glm based on the covariate structure. This can make it possible to assess goodness-of-fit in some models fitted to binary observations.

Usage

```
## S3 method for class 'glm'
group(object, eval = TRUE, ind = NULL, ...)
```

Arguments

`object`	a binomial glm object
`eval`	should the new glm-model be evaluated?
`ind`	an indicator for which rows to keep. If this is not specified the grouping structure is based on the covariate structure in the model.
`...`	currently not used

Details

The residual deviance and residual Pearson deviance are not meaningful measures of goodness-of-fit if the expected frequencies under the model are small (say less than five).

if `eval = TRUE` it is tested whether the estimated coefficients are identical up to three significant digits and a warning is issued if this is not the case. This should be the case in well-behaved situations but may not happen in cases of complete separation.

Value

A list with components

`newCall`	the new call
`newData`	a data frame with the aggregated data set
`oldData`	a data frame with the original data set
`oldN`	the number of rows (cases / observations) in the original data set
`newN`	the number of rows (cases / observations) in the aggregated data set
`oldObject`	the original fitted model
`newObject`	if `eval = TRUE` the new fitted model object, otherwise empty

Author(s)

Rune Haubo B Christensen

References

Collett, D. (2003) *Modelling binary data*. Second edition. Chapman & Hall/CRC.

Venables, W.N. and Ripley, B.D. (2002) *Modern Applied Statistics with S*. Fourth Edition. Springer

Examples

```
## Lifted from example(predict.glm):
ldose <- rep(0:5, 2)
numdead <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
sex <- factor(rep(c("M", "F"), c(6, 6)))
SF <- cbind(numdead, numalive=20-numdead)
## budworm.lg <- glm(SF ~ sex*ldose, family=binomial)
## summary(budworm.lg)
dat <- data.frame(SF=SF, sex, ldose)
dat[10, 1:2] <- rep(5, 2)
dat[13, ] <- dat[10, ]
rm(SF, sex, ldose)
SF <- as.matrix(dat[,1:2])
dat <- dat[,-(1:2)]
dat <- as.data.frame(cbind(SF, dat))

summary(m0 <- glm(SF ~ sex*ldose, binomial, dat))

## Various types of grouping:
(ind <- c(1:12, 10))
g <- group(m0, ind=ind, eval=TRUE)
g <- group(m0, eval=FALSE)
g <- group(m0, eval=TRUE)

## The correct GOF-test from the residual deviance is given by:
g$newObject
```

`halfnorm` *Half normal plot with simulated envelopes*

Description

`halfnorm` produces a half normal plot of the residuals with simulated envelopes useful for model evaluation and detection of outliers

Usage

```
halfnorm(object, resType = c("approx.deletion", "exact.deletion",
         "standard.deviance", "standard.pearson", "deviance",
         "pearson", "working", "response", "partial"), env = T,
         nsim = 20, plot = T, identify = F, n = 2)
```

Arguments

`object`	An object of class `glm` with a binomial family
`resType`	The type of residual used in the plot

env	Logical for whether envelopes are simulated
nsim	Number of simulations used for the envelopes
plot	Logical for whether the points should be plotted. If plot = F a list is returned
identify	Logical for whether it should be possible to identify points interactively. Ignored if plot = F
n	How many points should be identified. Ignored if identify = F

Details

Absolute values of the residuals are used in a half normal plot that otherwise corresponds to a regular normal probability plot.

Residuals from a binomial glm are not necessarily uncorrelated and normally distributed and may accordingly deviate from a straight line even if the fitted model is true. If the fitted model is true the optional simulated envelopes are likely to contain the absolute residuals.

The different types of residuals are described in Residuals

Value

If plot = T a plot is produced. Otherwise a list of the residuals and their expected values are returned

Author(s)

Merete K Hansen

References

Atkinson, A. C. (1981) Two graphical displays for outlying and influential observations in regression. *Biometrika*, **68**, 13-20.

Collett, D. (2003) *Modelling binary data*. Second edition. Chapman & Hall/CRC.

See Also

Residuals, identify

Examples

```
## Halfnormal plot with simulated envelopes
data(beetles)
beetles.glm <- glm(cbind(y, n-y) ~ conc, family=binomial, data=beetles)
halfnorm(beetles.glm, resType='pearson')

## Not run:
## Halfnormal plot with simulated envelopes
## Two points are interactively identified when they are selected with the mouse
halfnorm(beetles.glm, resType='deviance', identify = T, n = 2)

## End(Not run)
```

HLtest.Rsq *Goodness-of-fit tests for GLMs for binary data*

Description

Goodness-of-fit tests for GLMs for binary data including the Hosmer-Lemeshow decile test and X-squared test with normal approximation.

Usage

```
## S3 method for class 'Rsq'
HLtest(object, method = c("deciles", "fixed"),
         decile.type = 8, ...)

## S3 method for class 'HLtest.Rsq'
print(x, digits = getOption("digits"), ...)

## S3 method for class 'Rsq'
X2GOFtest(x, ...)

## S3 method for class 'X2GOFtest.Rsq'
print(x, ...)
```

Arguments

object	An Rsq object
x	An HLtest.Rsq or an X2GOFtest.Rsq object
method	The type of Hosmer-Lemeshow test to be performed. The "deciles" method should be more accurate (Hosmer et al, 1997)
decile.type	The quantile computation method; see quantile for details
digits	the desired number of printed digits
...	currently not used

Details

These tests are known to have very low power. They are only appropriate when the fitted frequencies are very low and when the covariate pattern dictates strictly binary observations.

Value

For HLtest.Rsq an object of class HLtest.Rsq with components

expected	the expected frequencies in the 2 x 10 entries
observed	the observed frequencies in the 2 x 10 entries
resid	Pearson residuals

`X2`	the Pearson X-squared statistic
`p.value`	the p-value for the goodness-of-fit test
`method`	the method used for the test

For `X2GOFtest` an object of class `X2GOFtest` with components

`p.value`	the p-value for the goodness-of-fit test
`z.score`	the standardized z-score for the goodness-of-fit test
`RSS`	the residual sums of squares term
`X2`	the pearson chi-squared statistic

Author(s)

Rune Haubo B Christensen

References

Hosmer, D.W. and Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics - Theory and Methods*, **A9**(10), p. 1043-1069.

Examples

```
## Lifted from example(predict.glm):
ldose <- rep(0:5, 2)
numdead <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
sex <- factor(rep(c("M", "F"), c(6, 6)))
SF <- cbind(numdead, numalive=20-numdead)
budworm.lg <- glm(SF ~ sex*ldose, family=binomial)
## summary(budworm.lg)

(Rsq.budworm <- Rsq(budworm.lg))

HLtest(Rsq.budworm)
HLtest(Rsq.budworm, method="fixed")
X2GOFtest(Rsq.budworm)
```

`profile` *Profile likelihoods for parameters in binomial regression models*

Description

Generate and plot the profile likelihoods for each parameter in a binomial regression model

Usage

```
## S3 method for class 'glm'
profile(fitted, which.par, alpha = 0.005, max.steps = 50,
         nsteps = 8, step.warn = 5, trace = F, ...)

## S3 method for class 'profile.glm'
plot(x, which.par, likelihood = TRUE,
         log = FALSE, relative = TRUE, approx = TRUE, conf.int = TRUE,
         level = 0.95, n = 100, fig = TRUE, ylim = NULL, ...)
```

Arguments

fitted	An object of class glm with a binomial family
x	An object of class profile.glm
which.par	A numeric or character vector with the parameters to be profiled. If missing all parameters are profiled
alpha	The likelihood is profiled in approximately the 100*(1-alpha)% confidence region
likelihood	Logical for whether the profile likelihood or likelihood root should be plotted
log	Logical for whether the profile likelihood should be plotted on log-scale. Ignored if likelihood = FALSE
relative	Logical for whether the profile likelihood or log-likelihood should be plotted on a relative or absolute scale. Ignored if likelihood = FALSE
approx	Logical for whether a quadratic approximation should be included in the plot
conf.int	Logical for whether a confidence interval should be included in the plot
level	A scalar or numerical vector indicating the confindence level(s) to be included in the plot. Ignored if conf.int = FALSE
n	How many points to employ in the spline interpolation of the profile likelihood
fig	Logical for whether the profile likelihood should be plotted. If fig = FALSE the list of points from the spline interpolation is returned
ylim	The limits of the y-axis in the plot
trace	Logical for whether progress should be printed to the screen during the profiling process
nsteps	Number of profiling steps to take in each direction for each parameter. The number is approximate since the step size is determined according to a quadratic approximation to the profile log-likelihood, hence, the deviation of the value of nsteps to the actual number of steps performed is influenced by the degree of irregularity of the profile likelihood
max.steps	The maximum number of profiling steps in each direction for each parameter. A warning is issued if the number of max.steps is reached
step.warn	A warning is issued if the the actual number of steps in either direction does not exceed the number of step.warn
...	Additional arguments passed to other methods

Details

`lroot` returned by `profile` is the signed square-root of the usual profile likelihood

$$sgn(\theta - \hat{\theta})\sqrt{2(l(\hat{\theta}) - l(\theta))}$$

where θ is the parameter being profiled and $\hat{\theta}$ is the maximum likelihood estimate of θ. The appertaining `par.vals` is a vector of θ values in an appropriate range around $\hat{\theta}$.

The logical argument `likelihood` in `plot` controls if the profile likelihood or the likelihood root should be plotted.

Value

For `profile`: a list of class profile.glm with a range of parameter values and lroot statistics for each parameter in `which.par`

For `plot`: if `fig = FALSE` a list with plotting points and confidence interval(s) for each parameter in `which.par` is returned. If `fig = TRUE` the list is returned invisibly.

Note

The implementation of these functions are largely inspired by `profile.glm` from the `MASS` package and `profile.clm` from the `ordinal` package. This work is a direct extension of `profile` from `MASS` with an extended set of warnings. The main difference, though, is in the plotting functionality, which enables plot of the usual profile likelihood and log-likelihood and the optional inclusion of confidence interval(s).

Author(s)

Merete K Hansen

References

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

See Also

`glm`, `profile.glm`, `plot.profile`

Examples

```
data(serum)
serum.glm <- glm(cbind(y, n-y) ~ dose, family=binomial, data=serum)
pr <- profile(serum.glm)
plot(pr)
```

`Residuals` *Residuals from a binomial regression model*

Description

Function to extract residuals from a binomial regression model

Usage

```
Residuals(object, type = c("approx.deletion", "exact.deletion",
          "standard.deviance", "standard.pearson", "deviance",
          "pearson", "working", "response", "partial"))
```

Arguments

`object`	An object of class `glm` with a binomial family
`type`	The type of residuals to be returned. Default is `approx.deletion` residuals

Details

A considerable terminology inconsistency regarding residuals is found in the litterature, especially concerning the adjectives *standardized* and *studentized*. Here, we use the term *standardized* about residuals divided by $\sqrt{(1-h_i)}$ and avoid the term *studentized* in favour of *deletion* to avoid confusion. See Hardin and Hilbe (2007) p. 52 for a short discussion of this topic.

The objective of `Residuals` is to enhance transparency of residuals of binomial regression models in R and to uniformise the terminology. With the exception of `exact.deletion` all residuals are extracted with a call to `rstudent`, `rstandard` and `residuals` from the `stats` package (see the description of the individual residuals below).

- `response`: response residuals

  $$y_i - \hat{y}_i$$

  The response residuals are also called raw residuals

  The residuals are extracted with a call to `residuals`.

- `pearson`: Pearson residuals

  $$X_i = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}}$$

  The residuals are extracted with a call to `residuals`.

- `standard.pearson`: standardized Pearson residuals

  $$r_{P,i} = \frac{X_i}{\sqrt{1-h_i}} = \frac{y_i + n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)(1-h_i)}}$$

  where X_i are the Pearson residuals and h_i are the hatvalues obtainable with `hatvalues`.

  The standardized Pearson residuals have many names including studentized Pearson residuals, standardized residuals, studentized residuals, internally studentized residuals.

  The residuals are extracted with a call to `rstandard`.

- `deviance`: deviance residual

  The deviance residuals are the signed square roots of the individual observations to the overall deviance

  $$d_i = sgn(y_i - \hat{y}_i)\sqrt{2y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + 2(n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)}$$

  The residuals are extracted with a call to `residuals`.

- `standard.deviance`: standardized deviance residuals

  $$r_{D,i} = \frac{d_i}{\sqrt{1 - h_i}}$$

  where d_i are the deviance residuals and h_i are the hatvalues that can be obtained with `hatvalues`.

  The standardized deviance residuals are also called studentized deviance residuals.

  The residuals are extracted with a call to `rstandard`.

- `approx.deletion`: approximate deletion residuals

  $$sgn(y_i - \hat{y}_i)\sqrt{h_i r_{P,i}^2 + (1 - h_i)r_{D,i}^2}$$

  where $r_{P,i}$ are the standardized Pearson residuals, $r_{D,i}$ are the standardized deviance residuals and h_i are the hatvalues that is obtained with `hatvalues` The approximate deletion residuals are approximations to the exact deletion residuals (see below) as suggested by Williams (1987).

  The approximate deletion residuals are called many different names in the litterature including likelihood residuals, studentized residuals, externally studentized residuals, deleted studentized residuals and jack-knife residuals.

  The residuals are extracted with a call to `rstudent`.

- `exact.deletion`: exact deletion residuals

  The ith deletion residual is calculated subtracting the deviances when fitting a linear logistic model to the full set of n observations and fitting the same model to a set of $n-1$ observations excluding the ith observation, for $i = 1, ..., n$. This gives rise to $n + 1$ fitting processes and may be computationally heavy for large data sets.

- `working`: working residuals

  The difference between the working response and the linear predictor at convergence

  $$r_{W,i} = (y_i - \hat{y}_i)\frac{\partial \hat{\eta}_i}{\partial \hat{\mu}_i}$$

  The residuals are extracted with a call to `residuals`.

- `partial`: partial residuals

  $$r_{W,i} + x_{ij}\hat{\beta}_j$$

  where $j = 1, ..., p$ and p is the number of predictors. x_{ij} is the ith observation of the jth predictor and $\hat{\beta}_j$ is the jth fitted coefficient.

  The residuals are useful for making partial residuals plots. They are extracted with a call to `residuals`

Value

A vector of residuals

Author(s)

Merete K Hansen

References

Collett, D. (2003) *Modelling binary data*. Second edition. Chapman & Hall/CRC.

Fox, J. (2002) *An R and S-Plus Companion to Applied Regression*. Sage Publ.

Hardin, J.W., Hilbe, J.M. (2007). *Generalized Linear Models and Extensions*. Second edition. Stata Press.

Williams, D. A. (1987) Generalized linear model diagnostics using the deviance and single case deletions. *Applied Statistics* **36**, 181-191.

Examples

```
data(serum)
serum.glm <- glm(cbind(y, n-y) ~ log(dose), family = binomial, data = serum)
Residuals(serum.glm, type='standard.deviance')
```

Rsq.glm *R-squared measures for binomial GLMs*

Description

This function computes the R-squared measures for binomial GLMs proposed by Tjur (2010) "Coefficients of determination in logistic regression models - a new proposal: The coefficient of discrimination".

Usage

```
## S3 method for class 'glm'
Rsq(object, ...)

## S3 method for class 'Rsq'
print(x, digits = getOption("digits"), ...)

## S3 method for class 'Rsq'
plot(x, which=c("hist", "ecdf", "ROC"), ...)
```

Arguments

`object`	a binomial `glm` object
`x`	an `Rsq` object
`which`	the desired plot: histograms, empirical cumulative distribution functions or ROC (receiver operating characteristic) curve
`digits`	the desired number of printed digits
`...`	currently not used

Details

The plot method has the following options

`"hist"` Two histograms with ten bins of the fitted probabilities are plottet on top of each other; the upper one for `y = 0` and the lower one for `y = 1`.

`"ecdf"` Two ecdf curves; one for `y = 0` and one for `y = 1`

`"ROC"` The (empirical) ROC curve

Value

`Rsq.glm` returns an object of class `Rsq`. The `plot` and `print` methods returns the `Rsq` objects invisibly.

Author(s)

Rune Haubo B Christensen

References

Tjur, T. (2009) Coefficients of determination in logistic regression models - a new proposal: The coefficient of discrimination. *The American Statistician*, **63**(4), 366-372.

See Also

A `HLtest` (Hosmer and Lemeshow test) method exists for `Rsq` objects.

Examples

```
## Lifted from example(predict.glm):
ldose <- rep(0:5, 2)
numdead <- c(1, 4, 9, 13, 18, 20, 0, 2, 6, 10, 12, 16)
sex <- factor(rep(c("M", "F"), c(6, 6)))
SF <- cbind(numdead, numalive=20-numdead)
budworm.lg <- glm(SF ~ sex*ldose, family=binomial)
## summary(budworm.lg)

(Rsq.budworm <- Rsq(budworm.lg))

plot(Rsq.budworm, "hist") ## or simply 'plot(Rsq.budworm)'
plot(Rsq.budworm, "ecdf")
plot(Rsq.budworm, "ROC")
```

`serum` *Anti-pneumococcus serum*

Description

Data from a study examining the protective effect of a serum co-administered in increasing doses with an infecting dose of a pneumococci culture. Each dose group consisted of 40 mice (n) and the number of deaths caused by pneumonia was recorded (y)

Usage

```
data(serum)
```

Format

A data frame with 5 observations on the following 3 variables.

`dose` Dose of the anti-pneumococcus serum administered

`y` Number of deaths in each dose group

`n` Total number of mice in each dose group

Source

Smith, W. (1932) The titration of antipneumococcus serum. *Journal of Pathology*, **35**, 509-526.

References

Collett, D. (2003) *Modelling binary data*. Second edition. Chapman & Hall/CRC.

Index